Universidad Autónoma de Madrid
Escuela Politécnica Superior
Facultad de Ciencias

Work to obtain master degrees in
ICT Research and Innovation (i^2-ICT) & Mathematics and Applications
by Universidad Autónoma de Madrid

Master thesis advisors:
Jorge E. López de Vergara Méndez - José R. Berrendero Díaz

Application of FDA to Network Management activities

David Muelas Recuenco

This work was presented on July, 2015

David Muelas Recuenco
Application of FDA to Network Management activities

David Muelas Recuenco
dav.muelas@uam.es

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

A ti

“ L’errore dell’intellettuale consiste [nel credere] che si possa sapere senza comprendere e specialmente senza sentire ed essere appassionato (non solo del sapere in sé, ma per l’oggetto del sapere) cioè che l’intellettuale possa essere tale (e non un puro pedante) se distinto e staccato dalla popolazione, cioè senza sentire le passioni elementari del popolo, comprendendole e quindi spiegandole e giustificandole nella determinata situazione storica, e collegandole dialetticamente alle leggi della storia, a una superiore concezione del mondo, scientificamente e coerentemente elaborata, il «sapere»; non si fa politica-storia senza questa passione, cioè senza questa connessione sentimentale tra intellettuali e popolo-nazione. ”

A. Gramsci

INDEX

PREFACE

Since I started the Joint Studies in Computer Science Engineering and Mathematics, I have realized the importance of exploiting the synergies between these two fields. During these (maybe few) years, I have seen how data analysis has been intensively widespread and included in many processes. I think that, with this extension, data analysis can help improving existent solutions to a huge variety of problems. Interestingly, not only scientific and technical, but also other problems related to many different areas —*e.g.* economics, social studies,...

As data analysis is a common space for computer scientists and mathematicians, I really feel that the dialogue between them is now more important than ever before. For me, the main reason to take the Joint Master's Degree in ICT Research and Innovation (i^2-ICT) and Mathematics and Applications (MMA) was the chance to improve this dialogue.

This document is my last step to finish this Master's degree, and I have tried to capture this motivation during all the pages that conform it. It is a summary of the different activities I have been carrying out during the last two years, which have been without any doubt the most intense years of my life. It groups all my (modest) contributions to the scientific and technical community, my (modest) contribution to the dialogue between computer scientists and mathematicians.

My first (modest) contribution to the improvement of the existent solutions to some problems. I hope it won't be the last one.

ACKNOWLEGMENTS

This section is intentionally in Spanish!

Sois muchas personas a las que tengo que agradecer haber llegado hasta aquí. Tantos que seguro que me dejo a alguien. Si es así, por favor entended y no me tengáis en cuenta la omisión.

Empiezo por los agradecimientos a todos los que estáis a mi alrededor en el ámbito profesional / académico. Gracias a mis dos tutores, Jorge y José Ramón. Sin vosotros, este trabajo no habría sido posible. Gracias a todas las personas del HPCN. A los miembros más *senior*: Javier A., Luis de Pedro, Sergio, Iván, Paco, Gustavo y Juancho, por compartir vuestra experiencia. Al resto de estudiantes con quien he compartido el C-113: Jaime G., Marco, Germán, Carlos, Rafa, José Fernando, Paula, Rubén,...

En particular, quiero plasmar unos agradecimientos especiales para algunos miembros presentes y pasados del HPCN. A los que seguís en la Uni, José Luis, Víctor M. y Javi R.; a los que siguen en el vecindario, Pedro; y a los que estáis un poco más lejos, Felipe, Pepelucho, Víctor L. Sin esos cafés, comidas, y cenas, sin duda yo no sería ni el profesional ni la persona que soy. Gracias de corazón.

Por otro lado, quiero agradecer la labor de mis profesores, acompañantes durante toda mi trayectoria como estudiante desde el colegio hasta aquí. Gracias también a mis compañer@s de clase, en especial a l@s de la Doble Titulación, Mates, e Informática. Gracias a Quique, Irene, Javi G., Dani y Eze, por ser compañer@s y amig@s. Gracias a David A., Víctor L. y Carlos del Cacho, por haberme ayudado a compatibilizar prácticas y trabajo. Sois unos genios. Gracias a l@s que no menciono explícitamente para no eternizar los agradecimientos, por compartir buenos y malos momentos.

Entrando en los agradecimientos puramente personales, quiero empezar por mi familia. A mi padre, Ramón, por ser maestro tanto en la escuela como fuera de ella. A mi madre, María, por ser la voz que ha estado a mi lado desde el principio. Habéis sido un apoyo incondicional que nunca me ha fallado. Sin vuestra presencia, no habría llegado hasta aquí. A mi hermano, Javier, porque mi vida no habría sido igual sin sus incordios :D. A mis abuelos, Chencho y Balbina, Antonio y Ermila, porque han sido y serán ejemplo de vida. A mis tí@s, y prim@s, siempre a un paso para cualquier cosa. Gracias también a Alfonso y a Juana, porque están aquí al lado, y son también familia.

Por último, pero no por ello menos importante, muchísimas gracias a ti, Cris. Gracias por aguantarme, por apoyarme, por acompañarme en este camino. Gracias por estar ahí, porque en resumidas cuentas, *haces que la vida se me vuelva de colores*.

Gracias a tod@s.

Abstract

Nowadays, the storage and processing capacities of current computer systems allow the use of diverse and massive data sources to extract knowledge that can guide managerial decision-making processes. Nevertheless, data analysis methods face some issues related to the characteristics of current data sources. In this work, we will focus in two aspects of current challenges in this area. On the one hand, we will propose some alternatives to transform heterogeneous data to a format that eases the exploration and study to network managers. On the other hand, we will look to the improvement of the scalability, extensibility and results obtained by using such solutions. Given that network monitoring is a critical labor, as a result of current systems' dependency on communication infrastructures, this Master Thesis provides a study of the applicability of *Functional Data Analysis* (FDA) to Network Traffic Monitoring and Analysis (NTMA) and network management. FDA is a branch of statistics that study infinite dimensional random variables, extending statistical properties to functional spaces. As several sources of network management data can be interpreted as functional data, our hypothesis is that the functional perspective may produce several advantages during the study and knowledge-acquisition phases of network analysis processes. This fact points to the improvement of several tools and methods that are used by managers and that can result insufficient with the current advances in network infrastructures. With this objective in mind, we present several proposals and some empirical results to illustrate the advantages of FDA when applied to network data mining.

Resumen

Actualmente, las capacidades de almacenamiento y de cómputo de los sistemas informáticos permiten el empleo de fuentes diversas y masivas de datos para extraer conocimiento de interés para la toma de decisiones de gestión. Esta situación plantea retos en el área del análisis de datos orientados a, por un lado, transformar los datos obtenidos de diversas fuentes a un formato que permita su exploración y estudio; y por otro lado, garantizar la escalabilidad de las soluciones, y su extensión y mejora continua. Dado que desde el punto de vista técnico, el control y evaluación de las prestaciones de las redes de ordenadores es un punto crítico por la dependencia que tienen los sistemas informáticos actuales en las infraestructuras de comunicaciones, en este Trabajo Fin de Máster se plantea el estudio de la aplicación de las técnicas de análisis estadístico de datos funcionales (*Functional Data Analysis*, FDA) al área de gestión de red. Estas técnicas consideran variables aleatorias de dimensión infinita, de modo que estudian propiedades estadísticas en espacios funcionales. Dado que muchos de los datos que son tenidos en cuenta durante las tareas involucradas en la gestión de red se acomodan a esta interpretación, se plantea como hipótesis que FDA puede proporciona ciertos avances en el estudio y extracción de conclusiones para guiar las decisiones de los agentes de gestión de red. Con el fin de estructurar este estudio, se presentará un catálogo de propuestas, presentando los resultados de su aplicación a la minería de datos de red.

1 INTRODUCTION

1.1 Network management, Data Mining and Big Data

The importance of data mining and analysis in many fields is increasing. Particularly, the applications to business intelligence and managerial areas are acquiring a notable significance, as they provide means to make data-based decisions [1]. As a result, these processes have become interdisciplinary activities: they encompass tasks from different areas (*e.g.* statistics, computer architecture,...) and the incorporation of domain-dependent knowledge. This character is additionally illustrated by the definition of curriculum profiles for experts in this area, as that recommended by the ACM (Association for Computing Machinery) [2]. In our case, we will focus in the applications of data mining to define network management actions. Particularly, we will study the application of some recent statistical advances to the analysis of different network data sources.

As some examples of the increasing attention and interest in the field we are studying, we can point to works such as [3–7]. In those, authors faced different steps in data mining and analysis to detect issues related to the different network Systems management functional areas —namely, those included in the FCAPS (Fault, Configuration, Accounting, Performance, Security) model [8].

Furthermore, data analysis has become into a complex set of techniques. We are in the *Big Data era*, and the growth of both available data and computing capacities is changing the approaches of management in all the areas [9]. Nevertheless, the incorporation of Big Data into managerial processes must be carefully faced. As stated in [10], it is necessary to define a structure for both data- and process-flow when facing data analysis activities. These data and process flow structures must take into account the characteristics of the systems under analysis, which calls for the definition of domain-dependent models to control such analytical processes.

Given that in we are coping with the analysis of network data using novel statistical approaches, we first want to locate our contributions inside this data analysis flow. To do

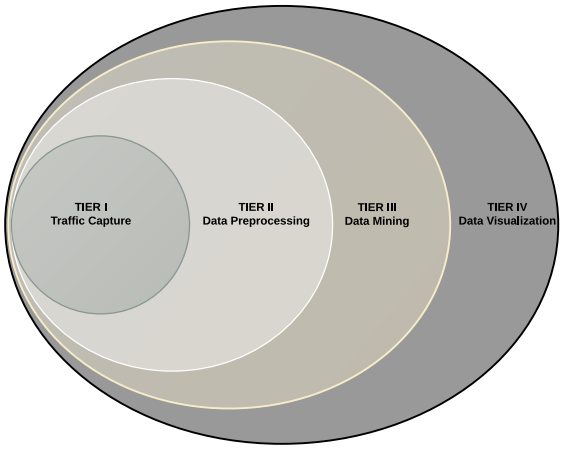

FIGURE 1.1: CONCEPTUAL MODEL OF THE TYPICAL STRUCTURE OF A NETWORK MANAGEMENT SYSTEM.

so, in Figure 1.1 we present a conceptual model for data analysis in network management activities. This model has been designed taking into account the characteristics of current solutions, which are composed of different functional elements that are related to the four tiers we distinguish:

- **Tier I: Traffic capture.** This tier includes all methods, activities and tools to get raw data from network infrastructures. We consider traffic as the basic data in network data analysis as it cannot be inferred from other data sources, and includes protocols that provide managerial data —*e.g.*, `IPFIX`, `NetFlow`, SNMP, server logs,...
- **Tier II: Data preprocessing.** Here we include the applications, frameworks and tools that use traffic to obtain other types of data —*e.g.*, network flow registers, different types of time series, statistical summaries,...
- **Tier III: Data mining.** This tier is composed by the methods, frameworks and tools that use traffic or derived network data to extract information and knowledge —*e.g.*, network modeling, pattern recognition,...
- **Tier IV: Data visualization.** It includes all functional elements that present analytical results to network managers.

We note that each tier includes the previous ones. This is a result of the usual access pattern of these systems: tier $n-1$ is the data source of tier n, and users interact with Tier IV. Additionally, tiers I to III correspond to the usual incorporation of data mining techniques in data acquisition processes in other domains. As an example, in [11], authors reviewed the effect of data mining in different databases, having in mind a cycle which is similar to our definition for the domain of networking.

In this context, we want to introduce some new concepts to the discussion of data

mining and analysis in the area of NTMA (Network Traffic Monitoring and Analysis). Thus, the advances that are presented in the following chapters are located in tiers III and IV of our model. The changing challenges that network management must face produce a continuous evolution of the suitable methods for network data analysis. Then, there is room for the improvement of current solutions affecting these tiers, which points to the exploration and exploitation of different techniques with the potential of surpassing previous solutions.

1.2 Main objectives

This work has been devised with the following **contributions** in mind:

- To describe the main results in the state of the art of two areas, namely NTMA and FDA (Functional Data Analysis).
- To evaluate different FDA techniques to represent and to analyze network measurement data.
- To open a new research line, which is based in the application of FDA techniques, that may be continued with the development of novel methods and advanced tools.

Having in mind these contributions, we have defined our main objective as **to study the applicability of FDA techniques to processes of data mining and analysis in the area of network management**. In this work, we will focus on the representation and mining of time series of network measurements and in the ECDF (Empirical Cumulative Distribution Function) of network flow characteristics.

We have additionally defined the following **partial objectives** to trace the evolution of our work:

- To study the state of the art of the areas related to the set of methods that are studied and proposed. This objective is covered in chapters 2 and 3.
- To summarize existent libraries to process FDA data, with their application in the domain of network management. This objective is covered in chapters 4 and 5.
- Evaluation of the effectiveness and limits of the considered FDA techniques in the domain of network management. This objective is also covered in chapters 4 and 5.
- Extraction of conclusions from the obtained results, and definition of a new research line in network management. This objective is covered in Chapter 6.

1.3 Structure of this document

To reach the objectives depicted in this chapter, the rest of this work is structured as follows. In Chapter 2 we review the main results of the state of the art of FDA. That chapter includes the description of the techniques that will be applied along our work to carry out several network data analysis tasks. To complete this discussion, in Chapter 3 we survey the current state of NTMA, summarizing the main methods, tools and systems that have been described in the literature of this area.

In chapters 4 and 5 we include the main findings and contributions of this work. We evaluate the application of several FDA techniques to the study of network measurements time series and to the study of network flow data, respectively.

Finally, in Chapter 6 we sum up the results of our work, extract and highlight some conclusions, and depict future work that conform a new research line in the area of networking.

2 Functional Data Analysis

2.1 Introduction

In this section, we survey the main results in FDA with focus on those that will be applied in the following chapters. Particularly, we describe techniques that illustrate the main differences and similarities with the case of multivariate analysis. This review is intended to provide a formal background for the following chapters, so in certain cases we will omit some technical and very theoretical aspects.

To do so, first we provide a description of the context of FDA in statistics. After that, we review the first steps when coping with FDA, and devote some sections to describe important results that will be applied in the following chapters. Finally, we highlight the conclusions that can be extracted from this chapter.

2.2 FDA in context

FDA is a set of techniques and statistical methods to analyze data belonging to infinite-dimensional spaces considering random variables which are themselves functions. This data interpretation can be suitable to study some problems, as data in many fields come to us through a process naturally described with a function —*e.g.* time series.

The application of FDA can provide different advantages. Roughly speaking, we can establish the following aims [12]:

- To represent the data in ways that aid further analysis.
- To display the data so as to highlight various characteristics.
- To study important sources of pattern and variation among the data.
- To explain variation in an outcome variable by using input variable information.
- To compare two or more sets of data with respect to certain types of variation.

Additionally, the consideration of functional data can be viewed with some historical perspective. In Table 2.1, which is extracted from [13], we represent the evolution of statistical approaches. Having this sketch in mind, we can informally say that the *"progress of the mathematical statistics can be described in terms of the conquest of new broader more sophisticated structures for $\mathcal{X}$ and Θ, in particular those corresponding to infinite-dimensional spaces"*, as stated by Cuevas in [13]. Regarding to FDA, we can consider that the starting point for all its further development was the Karhunen-Loeve expansion. This result was extensively applied to solve problems in the engineering field where the covariance structures of the involved process were known. After that, several works initiated the current approach to FDA during the late seventies. Since then, there has been a great development of functional data techniques that have been applied in a broad variety of subjects.

TABLE 2.1: HISTORICAL EVOLUTION OF STATISTICAL THEORIES.

Statistical theory	$\mathcal{X}$	Θ	**Dating back to**
Classical parametric inf.	$\mathcal{X} \subset \mathbb{R}$	$\Theta \subset \mathbb{R}$	1920s
Multivariate analysis	$\mathcal{X} \subset \mathbb{R}^d$, $n \gg d$	$\Theta \subset \mathbb{R}^k$, $n \gg k$	1940s
Nonparametrics	$\mathcal{X} \subset \mathbb{R}^d$, $n \gg d$	$\Theta \subset \mathcal{F}$	1960s
High dimensional problems	$\mathcal{X} \subset \mathbb{R}^d$, $n < d$	$\Theta \subset \mathbb{R}^k$	2000s
FDA	$\mathcal{X} \subset \mathcal{F}$	$\Theta \subset \mathcal{F}$	1990s

In the rest of this chapter, we will visit some of the elements that conform this relatively recent statistical approach. There are several works that survey the state-of-the-art of FDA. We will use the following ones as a guide to conform our review. The book by Ramsay and Silverman [12] is the main reference in this area, and we will use it to present the most classical FDA approaches and results. In [13], the author reviewed some results related to different aspects of the analysis of functional data, providing a more recent study of current techniques and possibilities in this field. Additionally, in [14], we can find an extensive description of the main FDA-method implementations, both for `R` and `MatLab`. Taking into account these references, we will present first a general description of several FDA aspects; and afterward we will further explain a selection of techniques that we will applied in chapters 4 and 5.

Before starting our discussion, we note that the practical approaches to FDA have been done from two main ways, namely *(i)* by extending multivariate techniques from vectors to curves and *(ii)* by descending from stochastic processes to empirical data. In this work, we follow the first approach as it results more natural because of our case-study data characteristics.

2.3 First steps and basic concepts

In this section, we will focus on some formal aspects related to the first steps when using FDA techniques, namely data representation and structure of the sample space; and on some general probabilistic results and their adaptation to the functional setup.

2.3.1 Data representation

In general, we can say that data can be considered as functional if at least one of the following conditions is met:

- A functional model is suitable to represent the considered random variables, as a result of the application of some operators —*e.g.* the relevant information is included in the derivative function.
- Data can be viewed (at least, theoretically) in a increasing finer grid which in the limit leads to a continuous-time observation.

In Figure 2.1 we show an example of functional data which is extracted from [15]. It corresponds to a sample of handwritten letters, and we have selected it to show the heterogeneity of the data that can be consider as functional. In chapters 4 and 5 we use other examples of functional data. In those chapters, we will use the FDA framework to study time series of network measurements and network flow parameters, respectively. With these examples, we will show *(i)* the advantages of considering time series as functional data; and *(ii)* the possibilities that the representation of scalar-random variable observations as functional data brings —we can use the CDF (Cumulative Distribution Function) to do so.

With this intuition about the type of data that can be considered as functional, let us describe the functional spaces where our observations live. Most theoretical developments require a real, separable Banach space as sample space $\mathcal{X}$ with norm $||\cdot||$. Nevertheless, very often it is necessary to have a Hilbert space structure for $\mathcal{X}$ —we recall that this structure provides an inner product $\langle\cdot,\cdot\rangle$. In the literature (see [13]), two standard elections are the Banach space of real continuous functions $x : \mathbb{T} \to \mathbb{R}$, with $\mathbb{T}$ a real compact interval, endowed with the supremum norm:

$$||x|| = \max_{t\in\mathbb{T}} |x(t)| \tag{2.1}$$

or the Hilbert space $L^2[\mathbb{T}]$, with $\mathbb{T}$ a real compact interval again and the usual inner product:

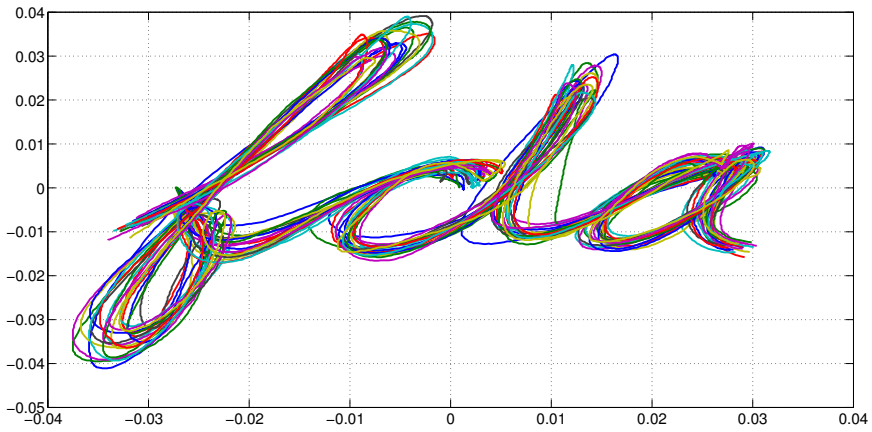

FIGURE 2.1: EXAMPLE OF FUNCTIONAL DATA. EXTRACTED FROM [15].

$$\langle x_1, x_2 \rangle = \int_{\mathbb{T}} x_1(t) x_2(t) dt \tag{2.2}$$

Other possibilities are available, depending on the characteristics of the data under analysis.

However, in empirical experiments it is not possible to obtain measurements in a continuous manner. Thus, the first step when using functional approaches is to interpolate and —if necessary— smooth the observations *with any global approximation technique*. In the literature, B-splines are a common election due to their properties [16], although other representations as, for example, Fourier series, are totally admissible if the structure of data is well preserved.

In general, we will represent the set of functions that conform the selected basis as

$$\{B_k(t)\}_{t \in \mathbb{T}, k \in \mathbb{Z}}$$

with $\mathbb{T}$ a real interval, and the coefficients giving the projections of the observations with respect the basis as $\{\beta_k\}_{k \in \mathbb{Z}}$. Thus, given a functional observation $\{X_t\}_{t \in \mathbb{T}}$ we can represent that observation as

$$\{X_t\} = \sum_{j \in \mathbb{Z}} \beta_j B_j(t), \ t \in \mathbb{T}$$

In practical applications, the representation is truncated so the representation would be given by the expression:

$$\{X_t\} = [\sum_{j\in\mathbb{J}} \beta_j B_j(t)] + \epsilon(\mathbb{J}, \{B_j\}),\ t \in \mathbb{T}$$

with $\mathbb{J}$ a finite set of indexes and ϵ a term of error dependent of both the set of indexes and the selected basis.

The previous concern about the use of *any global approximation technique* means that, during the estimation of the $\{\beta_j\}_{j\in\mathbb{J}}$ we will minimize the value of the quadratic error given by the expression:

$$\sum_{k=1}^{N}(X(t) - [\sum_{j\in\mathbb{J}} \beta_j B_j(t)])^2 = \sum_{k=1}^{N} \epsilon(\mathbb{J}, \{B_j\})^2 \tag{2.3}$$

Regarding data analysis practical issues, this representation presents several advantages:

1.– The amount of data required to describe the process is drastically reduced. The number of temporal points is usually much bigger than the number of components selected, and this fact leads to compact data representations.

2.– Robust estimations of the derivatives of the model can be obtained, as they can be explicitly calculated in terms of the basis expansion.

3.– It is possible to select the components containing the most relevant information about the model (*i.e.* PCA (Principal Component Analysis)). This representation enables to use dimensional reduction techniques based on the variance structure [17].

2.3.2 Some probability results

As a first insight to the FDA environment, we will study the adaptations of some probabilistic concepts to the functional setup.

Expectation

First, we consider the expectation, as it is natural to define and use this aspect in several inference processes. There are two natural ways to extend expectation to functional random variables:

- Using the usual process of the integral construction with a ascending hierarchy of complexity in the integrand.
- Defining a function by computing the usual expectation for each $t \in \mathbb{T}$.

We recall that, in general, we will consider random elements X defined on a Probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values in a Banach space $(\mathcal{X}, ||\cdot||)$. Thus, regarding the first procedure, we must construct the integral

$$\mathbb{E}(X) = \int_{\Omega} X d\mathbb{P} \tag{2.4}$$

using *(i)* Indicator function, *(ii)* Simple function and finally *(iii)* integrable functions, which can be defined as limit of simple functions. Using this first approach, we obtain the *Bochner integral* (also known as strong integral). Interestingly, the expectation of X exists if and only if $\mathbb{E}||X|| < \infty$, which gives an idea of the meaning of this definition [18].

On the other hand, the second definition of functional expectation that we are considering takes into account the fact that functional random variables are, in any sense, stochastic processes. Thus, if we interpret $X = X(t, \omega),\ t \in \mathbb{T},\ \omega \in \Omega$ we can define the function $\mathbb{E}(X) = \mathbb{E}(X)(t)$ as

$$\mathbb{E}(X) = \mathbb{E}(X)(t) = \int_{\Omega} X(t, \omega) d\mathbb{P}(\omega) \tag{2.5}$$

which leads to the *Pettis integral* (also weak integral). We should note at this point that, if the Bochner integral exists and it is finite, then it coincides with the Pettis integral. Moreover, the Pettis integral form for the functional expectation is close to the mean function of trajectories in the sample setup. We will use this second approach in chapters 4 and 5.

Mean, median and mode

Having in mind the definitions of functional expectation, we can use them to define the **mean** in terms of projections in the usual way. That is, the mean m of a random variable X with $\mathbb{E}||X||^2 < \infty$ fulfills:

$$\mathbb{E}||X - m||^2 = min_{a \in \mathcal{X}} \mathbb{E}||X - a||^2 \tag{2.6}$$

Regarding to **median**, we can obtain a functional definition in a similar way considering the minimizer of the function:

$$\mathbb{E}\left(||X - a|| - ||X||\right) \tag{2.7}$$

As an alternative, we can also use a depth-based adaptation of the median. We will further comment this in Section 2.5.

In the functional context, the absence of a unique natural notion of density makes it difficult to extend the notion of **mode**. Nevertheless, using a suitable kernel K we could define:

$$M_0 = \text{argmax}_a \mathbb{E}\left[K\left(\frac{||a - X||}{h}\right)\right] \tag{2.8}$$

which can be used in several data analysis processes. With this expression, we are defining an element of the sample space $\mathcal{X}$ which has a similar behavior to that of a mode. Nevertheless, there is no density to be approached when $h \to 0$ limiting the representativeness of this notion. This approach is quite similar to those related to the definition of principal curves and surfaces in the processes of manifold inference, learning and detection —*e.g.*, see the definitions in [19–21].

Needless to say, these concepts have their corresponding sample versions, which can be derived substituting expectations by the corresponding empirical averages.

Other results

There are other general probabilistic results that can be extended to the functional environment, such as large numbers' laws and central limit theorems. Nevertheless, those results are out of the scope of our study, so for the sake of brevity we will only mention their existence but we will no include further discussion.

2.4 Functional PCA

FPCA (Functional Principal Component Analysis) allows the selection of the projection directions that maximize the variance. PCA in FDA is conceptually quite similar to the corresponding classical technique, but it selects components that are combinations of the basis elements instead of selecting combination of different attributes. This characteristic provides a minor obfuscation of the semantic aspects of the resulting elements of representation, which is one of the main problems when applying PCA.

We recall that in the FDA context, instead of multivariate variable values we have function values $x_i(s)$. That is, the discrete index of each dimension of the multivariate variable is changed by a "continuous index" s. Taking into account the weights derived during PCA application, that we will denote with ξ, in the FPCA context we must adapt the discrete formulation to a continuous framework.

Then, the inner products:

$$\langle \xi, x \rangle = \sum_j \xi_j x_j \tag{2.9}$$

that appeared in the PCA definition for finite dimension vectors must be replaced by L^2 inner products, that is, integrals:

$$\int \xi x = \int \xi(s) x(s) ds \tag{2.10}$$

The weights ξ are now functions with values $\xi_j(s)$. The scores corresponding to each principal component are now given by the expression:

$$f_i = \int \xi x_i = \int \xi(s) x_i(s) ds \tag{2.11}$$

In the first FPCA step, the weight function ξ_1 (s) is chosen to maximize

$$\frac{\sum_i f_{i1}^2}{N} = \frac{\sum_i \int (\xi_1 x_i)^2}{N} \tag{2.12}$$

subject to the continuous analogue of the unit sum of squares constraint:

$$\int \xi_1(s)^2 ds = 1 \tag{2.13}$$

The following weight functions are required to satisfy the orthogonality restriction:

$$\int \xi_k \xi_m = 0, \ \forall \ k < m \tag{2.14}$$

Each function ξ_j define the most important mode of variation subject to the restrictions in equations 2.13 and 2.14. Note that the weight functions are defined only up to sign change.

This is the adaptation of the usual derivation of PCA to the functional context. Nevertheless, in the functional environment we can see the principal components as the basis functions that approximate the curve as closely as possible.

From this point of view, we can define the following functional expansion:

$$\hat{x}_i(t) = \sum_{k=1}^{K} f_{ik}\xi_k(t) \tag{2.15}$$

where f_{ik} is the principal component value $\int x_i\xi_k$. As a fitting criterion for an individual curve, we can consider the integrated squared error:

$$||x_i - \hat{x}_i||^2 = \int [x(s) - \hat{x}(s)]^2 ds \tag{2.16}$$

and as a global measure of approximation:

$$PCASSE = \sum_{i=1}^{N} ||x_i - \hat{x}_i||^2 \tag{2.17}$$

Thus, FPCA principal components are the basis functions that minimize those error functions. This is what motivates that, in some fields, functional principal components are referred as *empirical orthonormal functions*: because they are determined by the data they are used to expand.

In the functional version, the fact that the components that are selected are themselves "curves" imposes the necessity of some restrictions. To assure the representativeness of the components that are selected (*i.e.* to discard degenerated cases where they are too turbulent achieving the maximum variance), FPCA requires *(i)* the implantation of penalization elements in the optimization problem or *(ii)* to use the covariance function associated with smooth data.

There are other considerations regarding FPCA, related to extensions to bivariate and multivariate functions and computational considerations. In this work, we do not include this discussion, as it is out of the scope of our study. Nevertheless, a deeper review is offered in [12].

2.5 Functional depth

In 1975, John Tukey proposed a multivariate median, which is the "deepest" point in a given data cloud in $\mathbb{R}^d$. Since that moment, a rich statistical methodology based on data depth has been developed. General notions of data depth have been introduced as well

as many special ones. These notions vary regarding their computability and robustness and their sensitivity to reflect asymmetric shapes of the data. The notion of depth has been extended from data clouds to general probability distributions on $\mathbb{R}^d$ (which opened the gate to laws of large numbers and consistency results), and to functional data.

Depth measures in FDA are useful as they provide a notion of the *relative position* of elements in the set of observations. As several attempts to define and adapt L-statistics to functional data have appeared, depth measures have become a key element in constructing some statistics that require a certain order of the sample space.

The upper level sets of a depth statistic provide a family of set-valued statistics, named *depth-trimmed* or *central regions*, which describe the distribution regarding its location, scale and shape. In this sense, the most central region is equivalent to a median, the depth-trimmed regions can be seen as quantiles, and as a consequence, the depth of a data point is reversely related to its *outlyingness*.

Following the review in [22], we now describe some general characteristics of depth measures and a categorization based on the attributes that are used to quantify the centrality of observations.

2.5.1 General characteristics of depth measures

Let $\mathcal{X}$ be a Banach space, $\mathcal{B}$ its Borel set in $\mathcal{X}$, and $\mathcal{P}$ a set of probability distributions on $\mathcal{B}$. In the data analysis context, we may regard $\mathcal{P}$ as the class of defined by the ECDF.

A depth function is a function

$$\begin{array}{rcl} D : \mathcal{X} \times \mathcal{P} & \rightarrow & [0,1] \\ (z, P) & \rightarrow & D(z|P) \end{array} \tag{2.18}$$

with the following properties:

- **D1. Translation invariant.** $D(z + b|P + b) = D(z|P),\ \forall b \in \mathcal{X}$
- **D2. Linear invariant.** $D(Az|AP) = D(z|P)$ for every bijective linear transformation $A : \mathcal{X} \rightarrow \mathcal{X}$.
- **D3. Null at infinity.** $\lim_{z\rightarrow\infty} D(z|P) = 0$.
- **D4. Monotone on rays.** If an element z_0 has maximal depth, then for any r in the unit sphere of $\mathcal{X}$ the function $\alpha \rightarrow D(z_0 + \alpha r|P)$ decreases, in the weak sense, with $\alpha > 0$.
- **D5. Upper semicontinuous.** The upper level sets $D_\alpha(P) = z \in \mathcal{X} : D(z|P) \geq \alpha$ are closed $\forall\ \alpha$.

D1 and D2 state that depth functions are affine invariant. D3 and D4 mean that the level sets $D_\alpha, \alpha > 0$ are bounded and starshaped about z_0 . If there is a point of maximum depth, this depth will w.l.o.g. (without loss of generality) be set to 1. D5 is a useful technical restriction.

Additionally, we can follow this result from D4:

Theorem 1. *If P is centrally symmetric distributed about some $z_0 \in \mathcal{X}$, then any depth function $D(\cdot|P)$ is maximal at z_0.*

This result entails the maximization of functional depths in the median.

Taking into account the order induced by any depth function D, we can derive an *outlyingness function* given by

$$Out(z|P) = \frac{1}{D(z|p)} - 1 \tag{2.19}$$

which is 0 at the center and tends to infinite at *"infinity elements"*.

Now, we will focus in some depth proposals related to functional data. This depth measures allow to indicates how "deep" a function $z \in \mathcal{X}$ is located in a given finite cloud of functions $\in \mathcal{X}$.

2.5.2 Φ-depth

For $z \in \mathcal{X}$ and an empirical distribution P on $\{x_i\}_{i=1...n},\ x_i \in \mathcal{X}$ we can define a general functional data depth with the expression:

$$D(z|P) = \inf_{\varphi \in \Phi} D^d(\varphi(z)|\varphi(P)) \tag{2.20}$$

where D is a $d-$variate data depth satisfying the postulates D1 to D5. Here, $\Phi \subset E'^d$ and $\varphi(P)$ is the empirical distribution on $\{\varphi(x_i)\}_{i=1...n}$. We refer to D as Φ-depth. Thus, depending on the Φ that is used, we can obtain particular depth definitions —we address some of them below. Each φ can be viewed as a particular aspect of the functional data. Nevertheless, to hold the properties that are required to be a depth function, the family Φ must be carefully selected. Some relaxation on the postulates D1 to D5 can be contemplated so that the family Φ can be flexible enough to fit certain analytical processes. We will omit this formal discussion here, pointing to [22] for a deeper discussion.

2.5.3 Grid depths

Grid depths are based in the consideration of a finite dimensional tuple of values of trajectories with functions r such as $||r|| = 1$. These functions r act like weights, which allows to consider functional depths based on *(i)* projections (including some approaches related to principal components) and *(ii)* weighted averages.

2.5.4 Graph depths

To define this type of functional depths, we consider:

$$\Phi = \{\phi^t : \mathcal{X} \to \mathbb{R}^d : \phi^t(x) = (x_1(t), \ldots, x_d(t))\} \tag{2.21}$$

with t in the interval where the elements of $\mathcal{X}$ live and $x \in \mathcal{X}$.

As examples of this kind of functional depths in the literature, we indicate those included in [23, 24]. In particular, in the next chapters we will always consider the sample definition of the Half Region depth included in [24], given by the expression:

$$MS_{n,H}(x) = \min\{SL_n(x), IL_n(x)\} \tag{2.22}$$

where

$$\begin{aligned} SL_n(x) &= \tfrac{1}{n\lambda(\mathbb{T})} \sum_{t=1}^{n} \lambda\{t \in \mathbb{T} : x(t) \leq x_i(t)\} \\ IL_n(x) &= \tfrac{1}{n\lambda(\mathbb{T})} \sum_{t=1}^{n} \lambda\{t \in \mathbb{T} : x(t) \geq x_i(t)\} \end{aligned} \tag{2.23}$$

with λ the Lebesgue measure.

This definition is widely used, due to the low computational cost of this expression and its intuitive meaning. Roughly speaking, this functional depth is based on the fraction of "time" that the observation n is dominated and dominating other elements of the set of observations.

2.5.5 Multivariate depths

There are some proposals of multivariate functional depth [25, 26]. These notions provides the possibility to apply depth-based concepts not only for curves, but also for multivariate functions and surfaces. The use and study of these depth notions are part of the future work we plan, as we have restricted the scope of our discussion to the application of univariate functional depth notions.

2.6 Functional homogeneity

Given two samples, a natural question is when the observations from the two samples are realizations of the same stochastic process. In classical statistics there are many methods that can be used to test the homogeneity of two samples (*e.g.* χ^2 test).

In the field of FDA, there are some recent proposals to face this problem, providing some promising results. In [27], authors consider different homogeneity measurements based on the concept of functional depth. They use a computational approach to obtain an estimation of the distribution of the statistics that they are using to test if the functions come from the same model. They conclude, among other considerations about their method, that there is some homogeneity information in the derivatives of functional data, which calls as stated above to the joint consideration of functions and their derivatives. Additionally, in [28] authors propose some K-sample tests to study density functions that might be used to define homogeneity into this particular functional family.

Furthermore, we can regard to functional homogeneity as a concept related to the existence of more than a cluster into a set of functional observations. As we will comment in Chapter 4, this can be used as an approach to refuse homogeneity (with a multivariate analysis) if we use a set of components obtained after applying FPCA.

2.7 Phase-plane analysis

Phase-plane analysis describes the temporal evolution of a system making use of the relation between the value of a function and the associated value of its derivative. This representation of a system allows several analytic processes, such as stability studies for dynamical systems or the visualization of the evolution of a function value and its variation in a given point. As not only the pointwise value of a function but also its variation rate includes information about its analytical properties, this joint study improves certain data analysis —*e.g.*, homogeneity studies or clustering of curves.

If we consider a certain functional observation $f(t)$, we can obtain a phase-plane plot using *(a)* numerical estimations (*i.e.* with finite difference methods) or *(b)* analytical derivation using the functional representation previously described. Additionally, some smooth procedures and restriction can improve the latter approach, as commented in [29]. In that work, authors provide an analysis of the advantages of using multi-resolution analysis when estimating derivatives of a functional model.

2.8 Other FDA-based techniques

FDA includes other techniques, such as functional clustering, classification and forecasting Let us briefly comment some of them.

Supervised and unsupervised classification

In the literature, there are several works addressing functional classification, both supervised and unsupervised (clustering). As FDA techniques are constructed on Banach spaces or Hilbert spaces, it is possible to adapt the multivariate approaches by the use of the corresponding distances or inner products. Nevertheless, some formal issues arise, as those described in [13] —*e.g.*, those related to non-invertibility of the covariance operator or to the lack of universal consistency. Additional restrictions solve (at least partially) these matters.

It is also possible to map functional classification in the scope of the multivariate analysis, via the projection on finite-dimensional functional spaces and the study of the resulting coefficients. For example, in [30], authors present a supervised weighted distance approach, offering a complete comparison with other methods that use other projections. Other approaches, as that explained in [31], rely on depth-based projections to define discriminant functions, also when considering supervised classification.

Also it is possible to extend robust methods to the functional environment. In [32], authors proposed a method for unsupervised classification constructed on trimmed means. They claim that, with the restriction associated to this approach, it is possible to surpass problems related to outliers and deviated data.

Additionally, there are some proposals that use some proper characteristics of functional data. For example, in [33] authors described a clustering method based on a phase-amplitude consideration of the functional data. This work is particularly interesting, as in the functional scope sometimes data is not perfectly aligned —that is, not showing peaks and valleys at the precise same location. Approaches relying on these aspects of data do not require corrections by means of *warping*, which enhances results minimizing the process steps that must be accomplished.

Functional regression

This topic is beyond the objectives of this work, so for the sake of brevity, we will only briefly comment the existent approaches for functional regression. For further details in this field, we point to [12], where this matter is widely described.

First, if we consider linear regression, we can distinguish two main approaches depending on the type of response: specifically, functional regression with functional or scalar response. The construction of these models is based on the adaptation of the multivariate case, changing adequately the inner products in the finite-dimensional case by those corresponding to the particular functional space.

Regarding to non-parametric functional regression, we are looking for the optimal approximation of the response variable in terms of other functional random variables, with the typical definition in terms of the conditional expectation, which means that we must face an optimization problem adapted to the functional context.

2.9 Conclusions

In this chapter, we have reviewed some of the main results in FDA. We have presented the general methodology to obtain functional representations from empirically observed data, including in our discussion some technical and formal aspects regarding to the functional space where functional observations are defined. We have also presented some probability results that are key elements in a wide range of data analysis processes, describing how to adapt them to the functional environment. Furthermore, we have stated the structure of several methods that with a particular interest as they have proven to be applicable in many fields. With this review, we have stated the main similarities and differences between the multivariate statistical approaches and the corresponding ones in the case of infinite-dimensional spaces. As a result, we have provided a formal background for the discussion about the applications of functional methods to several activities in the NTMA domain. Regarding omissions in our review, to further study those aspects of FDA is devised as future work.

3

Network Management

3.1 Introduction

In this chapter, we review the current state of network management solutions. We will describe and comment the characteristics and shortcomings of current solutions, linking them to the novel approaches that we will describe in the next chapters. We have devised a review which is structured in terms of the conceptual model defined in Chapter 1 to provide a general insight of the context of our proposal.

To do so, we have structured the rest of this chapter as follows. First, we motivate the necessity of evolution of NTMA techniques as a consequence of the changes in network infrastructures. Next, we discuss the existent solutions in each of the tiers we have defined in Chapter 1. Finally, we describe a system architecture that incorporates some of the techniques described in Chapter 2 and that will be evaluated in chapters 4 and 5 and present the main conclusions extracted from this chapter.

3.2 Do we need further advances in network management?

Network management tasks are currently characterized by the diversity both in terms of the situations that must be faced and of the data used to extract conclusions. A huge amount of diverse data can be considered during the network management activities (*e.g.*, MRTG measurements, flow records or logs) providing information related to various layer issues in a wide range of network elements —*e.g.*, content server performance issues, misconfiguration, security issues,...

As a consequence, the management solutions applied in this area must evolve to accomplish the requirements that such context forces. Classical methods can result insufficient to researchers and practitioners if their hypothesis are not satisfied or if the ideal deployment scenarios are not in accordance with those under analysis. In this sense, the

appearance of changes on network dynamics is a potential source of erroneous results if the solutions lack of adaptive capabilities. For example, approaches that require the Gaussianity of network throughput are not adequate in situations where this condition is not met, restricting their applicability in some circumstances as several works study [34, 35].

Furthermore, the huge amount of data generated in modern computer networks that is processed and persisted during network management activities must be optimized to improve the scalability of the solutions. Hence, the selection and compression of informative records that fairly represent the network state is a problem of capital importance in network management activities, specially when dealing with long term analysis.

The encryption of the transmitted information and other legal and privacy aspects concerning this data limit some state-of-the-art solutions. As a particular situation that highlight the importance of this matter, if we focus on activities oriented to assure a certain security and anonymity level, intrusion detection systems that rely on DPI (Deep Packet Inspection) techniques can be totally useless.

Apart from the heterogeneous and complex contexts that must be considered, classical approaches do not take advantage of the capacities of processing huge amounts of data. In the literature, there is a huge variety of tools and methodologies to obtain different types of network measurements. Nevertheless, not only the measurements are important from the point of view of network management. Also the application of suitable techniques improves the quality and depth of the knowledge that can be extracted from measurements. Thus, once we have collected network measurements, managerial tasks require to extract conclusions from data using different data mining approaches.

The knowledge acquisition, which is necessary to reach conclusions, leads to a typical data and process flow that must be taken into account to obtain valuable findings when exploring network measurements. With this, the selection of the information that is used during certain tasks, as for example anomaly detection, can be enriched in the era of Big Data if we apply suitable analysis processes.

Unfortunately, monitoring systems provide network managers with tons of measurement data, and its interpretation has become a challenge. Using the conceptual description stated in [10], we can identify some steps to apply them during the study of network measurement data. First of all, it is necessary to extract general knowledge (*e.g.* models) from datasets containing the observations. Those models must provide meaningful information with high-level semantics that can help to understand the underlying phenomena. The application of this methodology also needs the consideration of privacy aspects, thus requiring sometimes additional obfuscation or deletion of some attributes —*e.g.*, user identifiers.

These are the shortcomings we try to surpass with our proposals. Our goal is to ease network managers' work by proposing novel approaches to study the behavior of network dynamics and flow characteristics. The term network dynamics refers to *the evolution*

of different macroscopic characteristics that define the network state. With network flow characteristic, we allude to *any metric that can be part of a typical or extended network flow record, as defined by IPFIX (IP Flow Information eXport)* [5] —some common examples are size in bytes, duration in seconds or number of packets. We will base our work in two different NTMA approaches.

First, network flow-based monitoring, which has received much attention by the research community as it represents a good trade-off between two opposite approaches, such as packet captures and aggregated time series —*e.g.*, MRTG outputs. This monitoring method has been proven useful to detect network intrusion, malfunction, or other types of anomalies. As an example, the authors in [36] show that under abnormal situations the size and duration of flows decrease at least one order of magnitude. Another example is that, during Denial of Service attacks, the proportion of flows with very few packets shoots up [37].

Second, the use of aggregated time series, which has also been deeply studied to detect Network traffic changes and abnormal patterns. In the beginning, the definition of static thresholds [38] (*e.g.*, if the ratio of small flows was over a certain value) was the typical approach to this problem. However, it is unable to provide the flexibility that monitoring requires. More recently, research efforts have been focused on applying statistics to this issue. First, only based on changes on mean and variance, and later, more complete studies based on histograms and cumulative distribution functions [39, 40].

To sum up, our proposals *(i)* extend current solutions to suit the changing operational characteristics of computer networks, and *(ii)* provide a manager-friendly output that eases interpretation and exploration of network measurements data. To do so, we will use FDA to explore network data measurements time series, as it provides a flexible framework to cope with observations from functional processes, and projections of functional summaries inferred from network flow observations, to summarize, homogenize and easily interpret per-flow data.

3.3 Current solutions and methods

In this section, we present related works that cover different tasks of capital importance in Network Management. Specifically, we describe solutions, tools and methods used to face different processes that conform some of the activities of each Systems management functional area of Network Management. To organize our review, we will use the conceptual description included in Chapter 1, grouping Tier I and II and Tier III and IV respectively. Additionally, in Figure 3.1 we provide a graphical summary of the elements we survey.

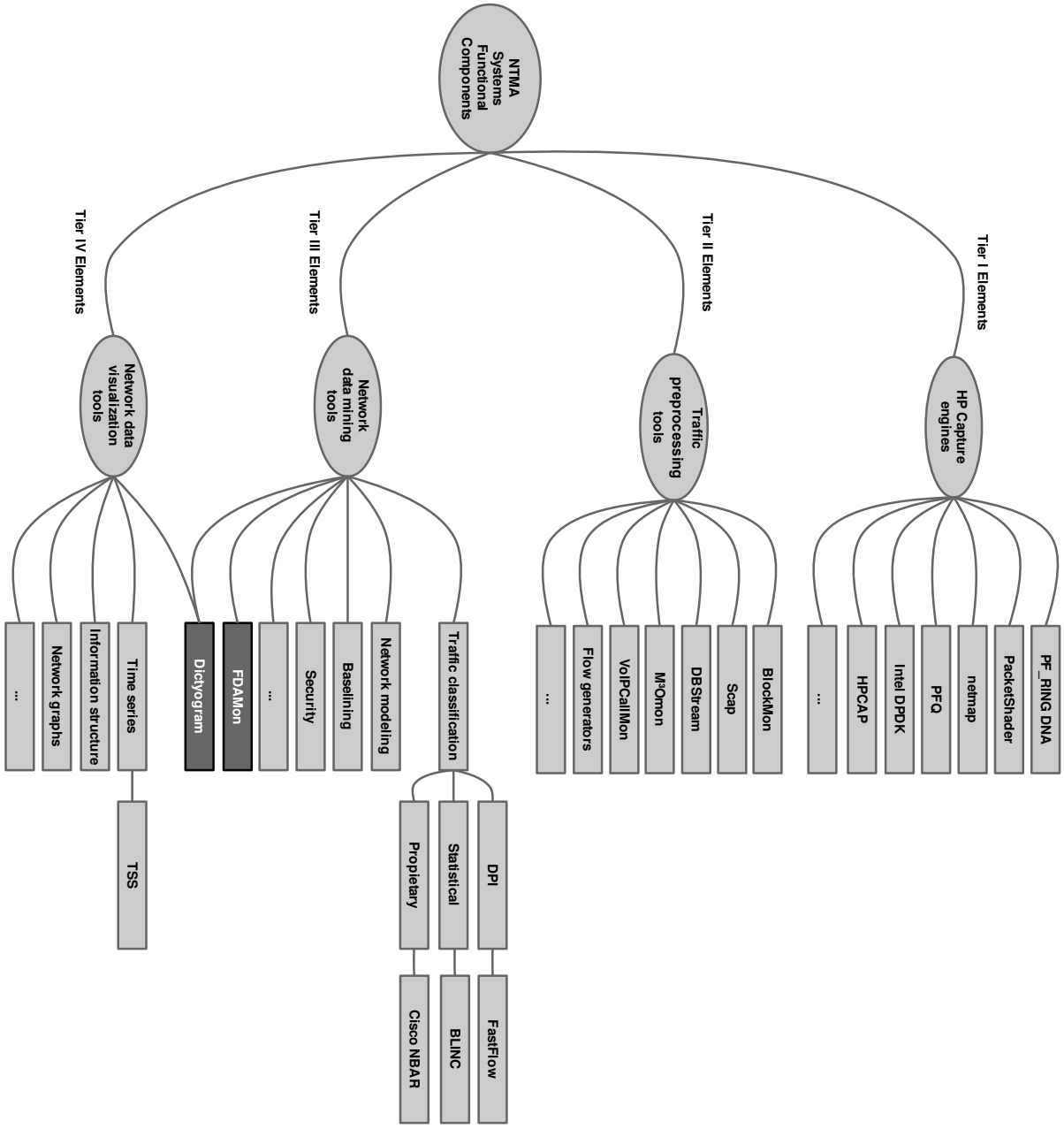

FIGURE 3.1: GRAPHICAL SUMMARY OF THE SURVEYED ELEMENTS.

3.3.1 Traffic capture and Network Data Preprocessing

In [41] authors reviewed current solutions related to high performance traffic capture engines. That work includes references to the main tools that have been proposed to cope with the challenging traffic rates that traverse multi-Gbps networks. Authors have evaluated different traffic capture engines (namely, they tested `PF_RING DNA`, `PacketShader`, `netmap`, `PFQ`, `Intel DPDK` and `HPCAP`) providing the results that have been used to select the Tier I elements in the analytical architecture that we present in the following section.

With respect to general network monitoring tools that provide a data source for our analysis by different means of traffic preprocessing (Tier II), we have selected some examples that illustrate the design principles of current proposals.

`BlockMon` is presented in [42], showing a modular and distributed monitoring framework which may be complemented with blocks implementing network analysis features. The authors indicate that future monitoring tools may follow this modular architecture in order to provide flexibility to developers and scalability by means of replication. This modular architecture makes possible the incorporation of Tier III modules with advanced functionality in the tools implemented using `BlockMon`.

Authors in [43] presented `Scap`, a complete framework oriented to stream-oriented analysis in demanding scenarios. `Scap` is in charge of traffic capture and stream reconstruction. This monitoring solution can be extended with data from Tier III in our model, in order to provide a tool with additional functionality – for example, real-time selection of traffic that must be persisted when detecting abnormal patterns in network behavior. `Scap` is a proof of the growing importance of aggregated statistics (*e.g.* SNMP data, network flow records) when monitoring multi-Gb networks as a result of the minimization of the processing load.

`DBStream` [44] is a framework that provides a framework constructed on `PostgreSQL` following the DSW (Data Stream Warehousing) paradigm. They claim that this system is able to carry out fast and flexible analysis across multiple heterogeneous data sources, and their tests suggest that their systems outperforms MapReduce systems in several analysis. As this is mainly a data warehousing solution, it could be used as a data source for the modules we are presenting.

Following with MapReduce and the associated data warehousing solutions in the network management domain, in [45, 46] authors proposed NTMA systems constructed on Hadoop. The main problem with their solution is that, as stated [44], Hadoop is not able to provide low latency access to data. Nevertheless, this approach allows to scale up long term analysis, and the inclusion of our solutions in the framework that is described in those works could be straightforward given the high compatibility between the selected implementation design of our proposal and the Hadoop ecosystem.

To conclude our review, we want to point to other systems which provide low latency traffic processing and generate flow records. In [47], authors present M^{3}Omon, which is a software layer that is able to elaborate derived data (namely, SNMP (Simple Network Management Protocol) time series and IPFIX records) from network packets. This framework suits into our proposal, and as presented below, it will be the data source for the proposed architecture. Similarly, authors in [48] propose VoIPCallMon, a QoS (Quality of Service) and QoE (Quality of Experience) analysis system for VoIP (Voice over IP) deployments. VoIPCallMon provides both signaling and multimedia call records, including several quality parameters. Thus, this tool could be extended with the functional elements of tiers III and IV that we are proposing, to define a versatile VoIP quality management system.

3.3.2 Network data mining and visualization

Regarding to network performance measurement, in [49] authors propose a new metric with reduced computational cost that condenses significant information when applied to Data Center monitoring. This approach highlights some of the principles included in our solution, but is restricted to a particular context. One of the advantages of FDA is that few *a priori* assumptions are made, so it can be widespread to almost every scenario.

Other approaches have also proposed statistical-sound mechanisms to characterize traffic but paying attention to macroscopic behavioral aspects of computer networks [35, 37, 50, 51].

Authors in [35] propose the use of α-stable distributions to model traffic in low aggregation points (*i.e.* small networks). Additionally, the deviation of some of the characteristic parameters labeled as normal are used to detect anomalies. The main problem with their proposal is the high computational requirements for the estimation of the parameters that define a particular element of this family of distributions. Taking into account the principles stated in [49], the deployment of this approach may be unfeasible in many contexts. With the lightweight versions of functional metrics, these problems can be avoided.

On the other hand, in [50, 52] authors present statistical network models using Gaussian processes. Particularly, the solution proposed in [50] is oriented to link capacity planning inside a network by inference on the busy hour, which restricts this kind of solutions to scenarios where real-time analysis of network state is not required. In [52], the described methodology is oriented to the detection of sustained changes in load utilization. The Gaussianity of traffic load is the base of these and other models, but it is an hypothesis that cannot be directly assumed in general [34, 35], as we mentioned before. FDA techniques do not suffer from this problem, as they no require assumptions relative to the marginal distribution of the considered parameters.

Let us now focus on previous works that studied preprocessing techniques for network

data mining. Authors in [53] propose the application of Principal Components Analysis (PCA) to throughput registers in order to decompose them in terms of *eigenflows*. That solution shares some similarities with functional PCA, which will be introduced in the following sections, but instead, it does not make use of a primary common representation of flows in terms of any type of components. As we will show, the advantage of having such previous representation is that there is no semantic obfuscation of data, which is one of the main problems when using PCA. Nevertheless, the central ideas of that work pinpoint to the gaining that a functional treatment of network parameters entails.

Other approaches that study data compression apply multiresolution analysis combined with the evaluation of statistical properties of the compressed data [54, 55]. Formally, the application of multiresolution analysis provides a functional representation of the data restricted to certain basis. As we will explain, this is the first step when applying FDA techniques, including among others the particular cases studied in these works.

Although the idea of using functional random variables that are defined in infinite dimensional spaces seems to be self-defeating, it is common to FDA and machine learning techniques that take advantage of the simplification of separation of sets when the dimension increases. For instance, Support Vector Machines (SVM) are a well-known example that has been successfully applied to diverse problems related to Network Management activities, achieving good results both in terms of efficiency and presision. In [56, 57], authors explore the results that SVMs provide in anomaly detection, management of Quality of Experiences (QoE) and QoE prediction.

Other methods directly oriented to the discrimination of anomalous behavior are described in [37, 51]. These approaches use Network Behavior Analysis (NBA) techniques to detect and classify patterns that might indicate the presence of any type of anomaly. NBA can be seen from the point of view of FDA as a set of functions that describe the network state, providing formal soundness to the analysis and a base to use all the advanced features that FDA encompass.

Nevertheless, the study of aggregates is often insufficient for certain situations as stated in [47] and usually flows represent a better trade-off between burden and precision. The authors in [58] provided an extensive review of current applications based on the concept of network flows. Such review includes several applications such as performance evaluation, misuse of bandwidth, and monitoring for QoS among others, between which traffic characterization, diagnostic, security and intrusion detection stand out. Our approach may be included in these latter categories, where we share space with works like [39, 40, 59].

In [39], authors found that flows can be categorized according to their size and duration in four categories, named by analogy as dragonflies (short), tortoises (long), mice (light) and elephants (heavy). Furthermore, the authors in [40] continue targeting traffic classification by flow size and duration, and define other classes such as the buffaloes, which are more spiky flows. They refine flow classification by using histograms and mod-

eling them with Dirichlet random distributions and a stochastic version of the Expectation Maximization algorithm. We note that these categories are intended to describe flow behaviors, not working as real mechanism to detect deviations from normal operation. In addition, we are proposing a simpler method to categorize flow characteristics, as well as a visual framework to show when the network went wrong.

Regarding Intrusion Detection Systems (IDS), the authors in [59] reviewed solutions based on the construction of IP flows. They provided a deep insight into the different approaches to identify problems in a network using flows. Among these methods, it is remarkable the proportion of ICMP flow, size and distribution of IP ranges, number of SYN packets and the number of SYN/ACKs, small ratio flow-size/packets among others. This illustrates the diversity of characteristics that our approach could exploit to detect network issues.

Let us now present some surveys on frameworks for network data visualization. In [3], the authors provided a review of existent systems oriented to detect security issues. That work analyzed different aspects of such systems, including data sources and classification criteria. Their conclusions pinpointed to the necessity of techniques that exploit the capacities of a human analyst when defining network data visualizations. In this sense, our proposal provides a "manager friendly" summary of the evolution of flow dimensions, without saturating them with irrelevant information.

The authors in [6] showed a framework based on both data mining and visual graphical representation. They proposed the integration of different tools in a unique integrated network traffic visualization system, and presented a number of examples. Our solution is a complementary visualization tool that can be included in such a framework, to provide a temporal visual description of changes in flow parameters.

Finally, we mention the tool `Time Series Solver` (TSS) [60], which is a tool for the analysis of time series based on network flow monitoring. TSS includes a battery of tests to apply on the time series data as those our solution outputs. However, they do not uniformly classify the flows, with the lack of semantic these classes provide.

3.4 Opportunities in Network Management

Keeping in mind the limitations of existent solutions, we describe a first overview of the benefits of the functional treatment of network management data. The – at least – theoretical possibility of observing network measurements time series at any time (*i.e.* as a continuous process) allows the consideration of them as functional data [13]. Furthermore, other network management data, as network flow characteristics, can be included into the functional world if we consider their ECDF. The good results of the application of **FDA** to face other problems that deal with data with this property, such as weather forecasting,

human growth and some economical studies[1], motivates the study of the applicability of FDA in the area of network management, as it is a branch of statistics offering a wide range of opportunities. For instance, if we focus on the temporal evolution of the parameters that define the network state, the consideration of the observed time series as realizations of a certain functional random variable holds a novel vision with peculiarities that, as we will illustrate, make some network monitoring tasks easier.

Thus, we will explore some FDA results in the NTMA area. On the one hand, we will consider the application of FDA to network data time series. Our analysis will be centered in the advantages concerning compression and representation of data, attending to the semantic and meaning of the transformations; in the description of multivariate techniques that provide a joint treatment of certain parameters, which results essential in many situations; and in the definition of novel metrics and methods with low computational cost, that can lead to robust interpretation of network dynamics. We will briefly point to other aspects of FDA techniques oriented to forecasting, clustering and classification of network traffic, in order to conform a global vision of the opportunities that this field of statistics offers to network researchers and practitioners. On the other hand, we will consider other applications of FDA in the field of network data analysis. Specifically, we will analyze the use of FDA elements to provide CDF models of network flow parameters.

These analytical elements are included in an experimental framework which is represented in Figure 3.2. Given their particular characteristics, we have selected HPCAP and M^{3}Omon as Tier I and Tier II elements. We have defined the other elements, namely `FDAMon` and `Dictyogram`, which are implemented in `MatLab`, `Python`, and `R`, as these languages have proven to be versatile, and can be easily integrated in Hadoop-based systems, which can be interesting for further developments and versions of this framework. Additionally, these languages include the following libraries:

- `fda` [15], implemented for `R`.
- `fda.usc` [61], implemented for `R`.
- `fdaMatlab` [2], implemented for `MatLab`.
- B-spline toolbox [3], implemented for `MatLab`.

Theses libraries include implementations of several FDA methods, and a huge variety of other numerical analysis elements. We plan to migrate the `MatLab` elements to implementations in any of the other two alternatives to provide a open-source and free alternative.

To show the usefulness of these novel solutions, we will consider different case studies in which we will analyze real traffic records from the academic network of Spain. Thus, for

[1] `http://www.psych.mcgill.ca/misc/fda/examples.html`

[2] `http://www.psych.mcgill.ca/misc/fda/downloads/FDAfuns/Matlab/`

[3] `http://www.mathworks.com/matlabcentral/fileexchange/27047-b-spline-tools/content/bspline_tools_1_2/doc/html/BsplineDoc.html`

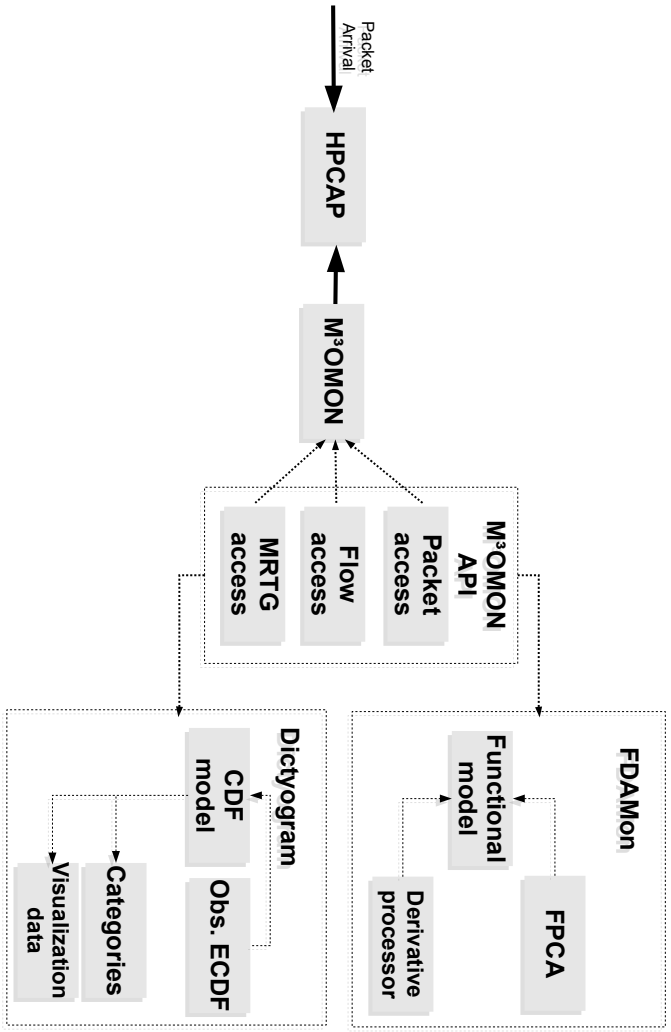

FIGURE 3.2: SYSTEM ARCHITECTURE FOCUSING ON THE FIRST THREE TIERS OF OUR CONCEPTUAL MODEL.

this evaluation, we will use network flow records persisted in text files.

3.5 Conclusions

Our survey has described current solutions in the different tiers that composed typical NTMA systems. With this review, we have stated the characteristics of several functional components for such systems, indicating the main limits and challenges in this area. From this study, we have found that the evolution of network analysis and management solutions must follow two main lines. On the one hand, future tools must provide means to cope with data heterogeneity and challenging generation rates. On the other hand, they must provide results in *(i)* a manager-friendly representation that eases the interpretation of data, and *(ii)* a rich manner that enables to reach meaningful conclusions and extract valuable knowledge. To do so, in this work we propose the application of FDA techniques, as the use of them could open the gate to advances in both lines. We have additionally presented an experimental framework that will be evaluated in the following chapters, that implement some methods with promising results.

4

Network Measurements data and FDA

4.1 Introduction

In this chapter we explore the advantages of FDA when applied to several network management and NTMA tasks. Specifically, we address the application of several functional methods to network measurement time series, describing the advantages of the functional approach. We show that, thanks to the peculiarities of the FDA techniques, it is possible:

- To obtain novel network-state representations with good resolution and reduced storage necessities.
- To provide data visualizations making use of functional aspects —*e.g.*, phase-plane plots.
- To estimate fine-grained baselines.
- To define flexible and extensible anomaly detection methods.

Additionally, functional methods open the gate to a novel interpretation of anomalous events thanks to functional clustering and classification.

To address these objectives, first, we describe the representation of network measurement data in terms of functional elements. Then, we study FPCA as a data mining preprocessing step for network measurements. After that, we illustrate some of the techniques described in Chapter 2, obtaining promising results that could provide a next step in network management activities. Finally, we present some conclusions that can be extracted from this chapter.

4.2 Network Management data representation

FDA allows the development of compact expressions of network parameters (packets, flows, bytes, active IP directions,...) represented as a function of a certain set of param-

eters —*e.g.*, time series, if they are represented as functions of time. This is particularly interesting when defining baselines [4], as it provides a continuous time approximation.

Additionally, if we define a function for the representation of the network state as $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we are using surfaces or curves describing the joint behavior of an arbitrary number of parameters.The joint consideration of several parameters is necessary to detect the appearance of some types of anomalies (*e.g.* some DDoS (Distributed Denial of Service) attacks [47]), which makes this feature of great utility for network managers.

Figure 4.1 shows the result of throughput data interpolation using third grade B-splines. This representation is obtained using a set of sampled points. After having obtained that functional expression, we evaluate the resulting curve for each time point. To obtain these results, we have used the B-spline toolbox for `MatLab`. At the same time, Figure 4.2 shows an extended set of throughput observations (546 days with a 5-minutes granularity) treated with the package `fda` of `R` [61]. In this case we have not used any sampling to obtain the functional expression, as this approach suits the analysis we carry out below. Both sets have been obtained from the Spanish academic network.

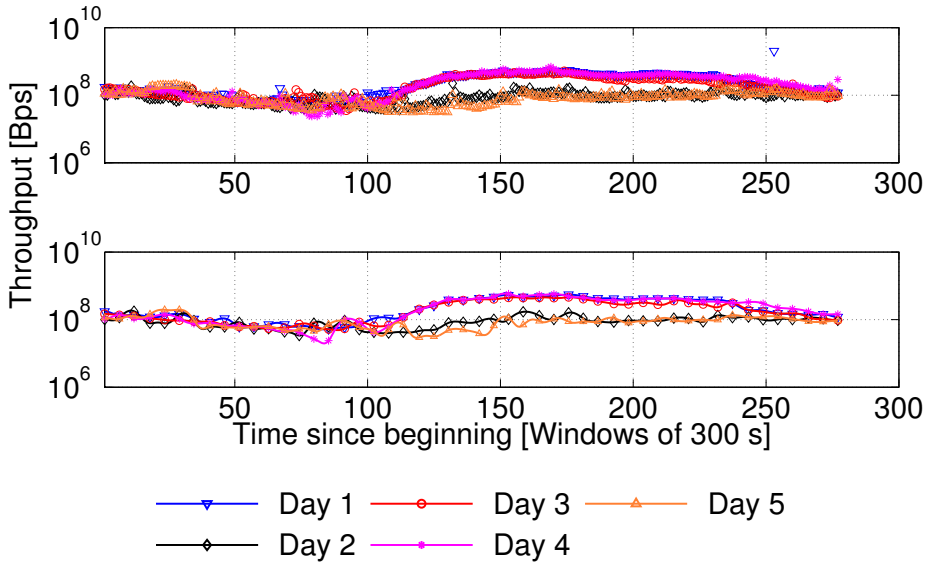

FIGURE 4.1: THIRD GRADE B-SPLINES REPRESENTATION FOR 5 DAYS OF THROUGHPUT REGISTERS, SPANISH ACADEMIC NETWORK.

The representation of network parameters as functional elements entails a first-level compression, as once we have selected a certain basis we only have to consider the coefficients that represent those particular elements. This aspect of the application of FDA to Network Management tasks results interesting when we consider scalability issues, at the same time that provides a first step to apply other FDA techniques that will be studied in the following subsections.

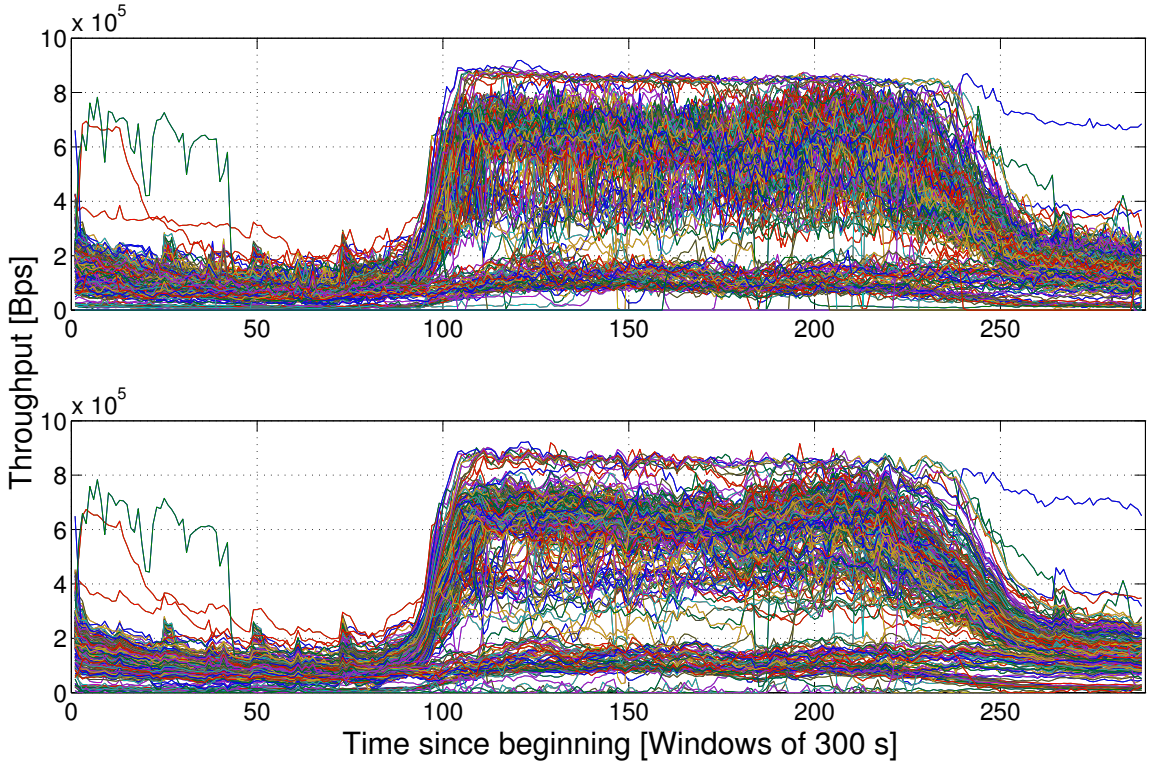

FIGURE 4.2: THIRD GRADE B-SPLINES REPRESENTATION FOR 546 DAYS OF THROUGHPUT REGISTERS, SPANISH ACADEMIC NETWORK.

4.3 FPCA for network management data

The application of FPCA to the representation of network management data entails several advantages. As previously mentioned, this technique selects combinations of basis elements in terms of the proportion of the original variance that they cover. This dimensionality reduction provides, on the one hand, a mean of data compression maintaining the original variability structure. On the other hand, the harmonics which are the output of FPCA describe curves that represent certain aspects of the observations with an interpretation in terms of the variability structure.

To illustrate these ideas, we apply FPCA to an extended set of throughput registers of the Spanish academic network, consisting of 20 daily observations interpolated with a third-degree B-spline basis of 60 components. The computation of FPCA on this set of observations has been obtained using the package `fda` of `R` [61]. Figure 4.3 represents the first 6 harmonics —that is, the number of components needed to cover the 95% of the original variance. With respect to the compression, the result allows the number of needed to represent the data to be reduced by a factor of ten. With respect to the aspects represented by each harmonic, notice that the first principal component, highlighted in the figure, represents a scaled *approximation* of the dynamic of the observations —omitted for the shake of clarity, given that they are similar to those represented in Figure 4.1. With the consideration of additional principal components, we enrich the representation with *details* that cover a higher proportion of the observed variability.

As a result of these properties, this method allows the reduction of the volume of the data that must be persisted with a criterion based on the variability structure. Compared to other alternatives as those commented in Section 3.3, FPCA harmonics represent a meaningful decomposition of the observations, instead of only a filtered or reduced version of them.

At the same time, in Figure 4.4 we show the results for the other considered dataset, taking into account in this case 30 components. As we have not applied any smoothing process when obtaining the functional representation of such data, we can see that the reconstructed and original time series have similar visual behavior. Additionally, in Figure 4.5 we represent coefficient densities for each component. It is worth remarking that, with this approach, it is straightforward to detect two clusters inside the set of curves.

FPCA provides several advantages in problems derived of certain network management activities. First of all, it provides a second-level compression of data, as we can select a subset of the principal components controlling variability information losses. Furthermore, changes in a certain harmonic indicate different types of changes in the measured parameter depending on the variability covered by such harmonic. This fact motivates a novel vision of anomalies and other variations of network dynamics, linking them to the behavior of the functional principal components. Finally, the semantic information that this decomposition entails allow to define sensitive baselines in terms of the evolution

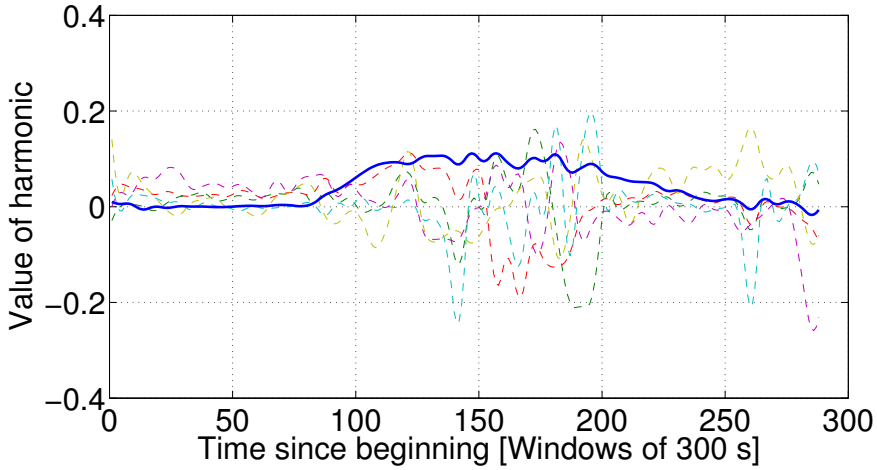

FIGURE 4.3: HARMONICS COVERING 95 OF THE ORIGINAL VARIANCE (6 COMPONENTS SELECTED) AFTER APPLYING FPCA ON THROUGHPUT RECORDS, ACADEMIC NETWORK. THE FIRST COMPONENT IS HIGHLIGHTED.

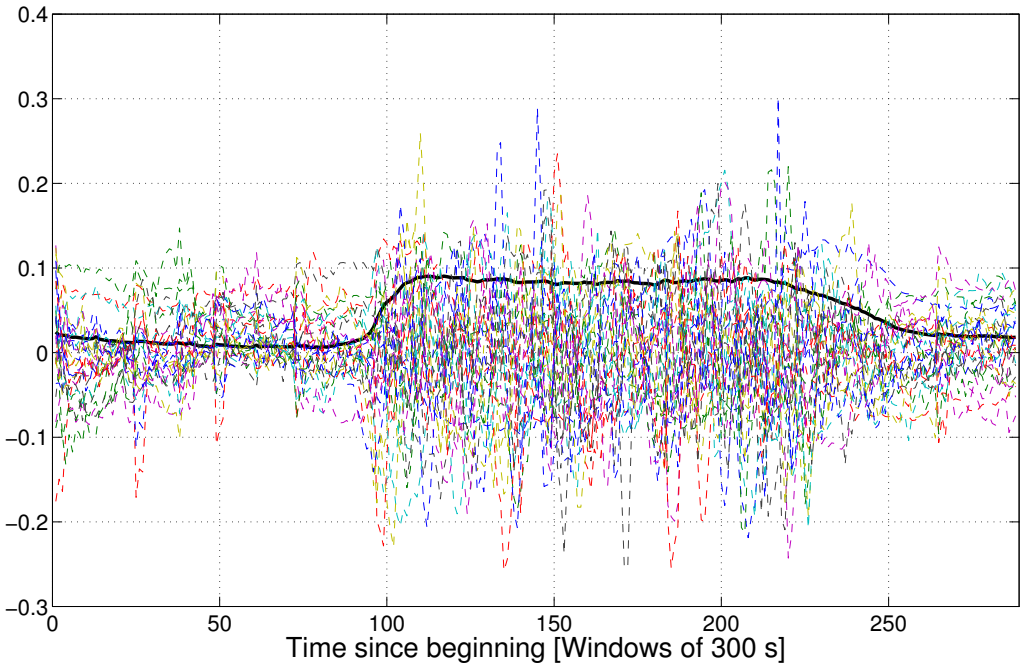

FIGURE 4.4: FIRST 30 HARMONICS, ACADEMIC NETWORK. THE FIRST COMPONENT IS HIGHLIGHTED.

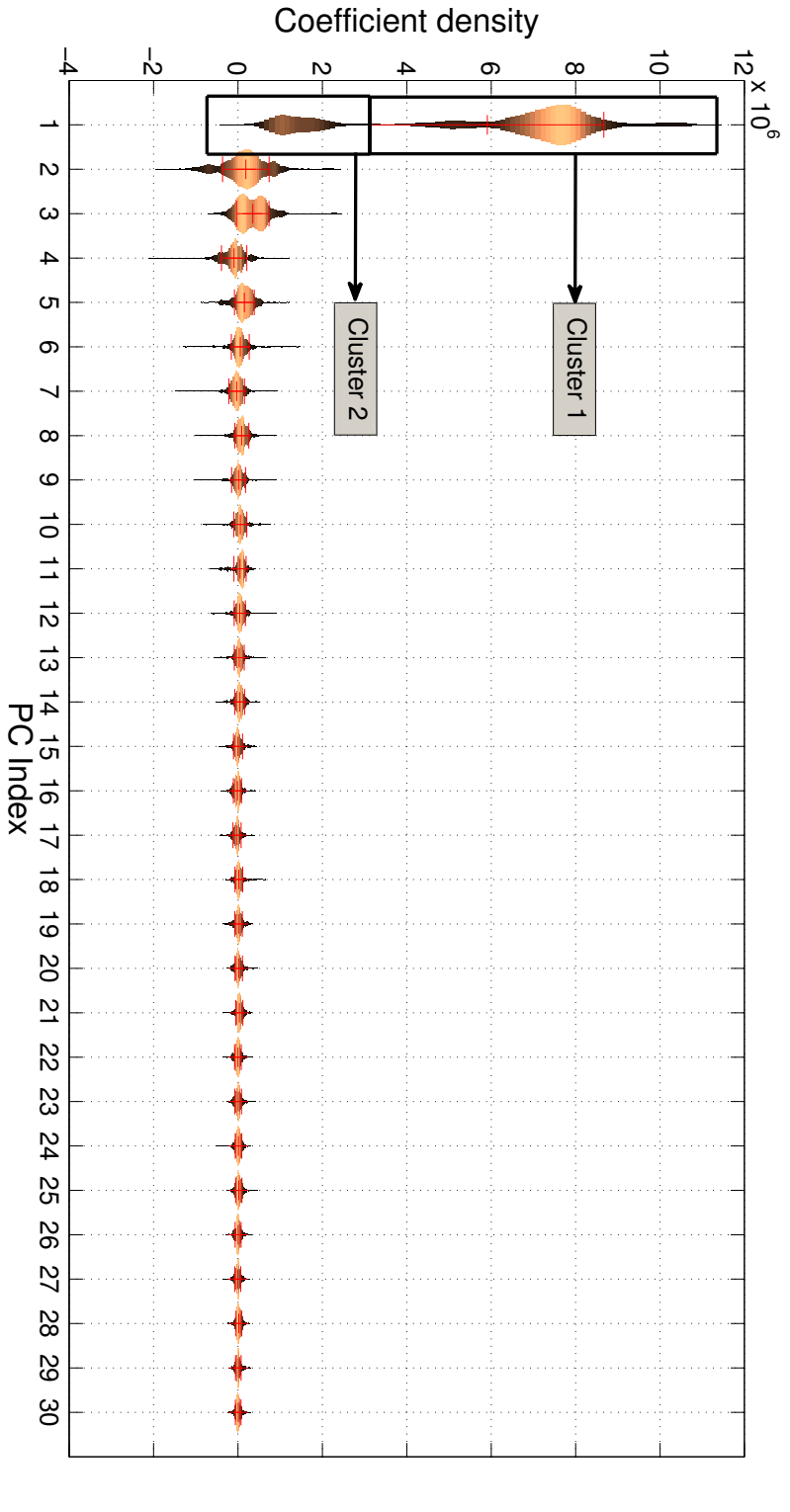

FIGURE 4.5: DENSITY OF COEFFICIENTS IN FPCA REPRESENTATION, EDUCATIONAL NETWORK. TWO CLUSTERS CAN BE DISTINGUISHED.

of the harmonics.

4.4 Phase-plane analysis for network management data

As mentioned in the Section 2.7, phase-plane analysis is a technique which links the behaviors of a function and its derivative. Since not only the value of parameters, but also change speed is important in several tasks of Network Management, this is an approach that can be useful in various processes of decision making.

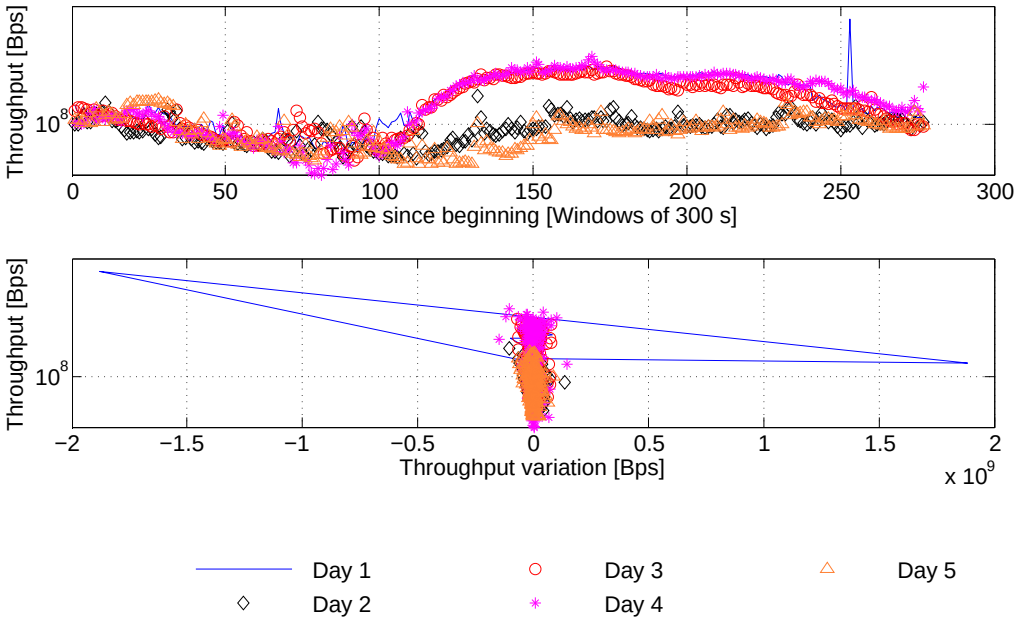

FIGURE 4.6: PHASE-PLANE PLOT FOR 5 DAYS OF THROUGHPUT REGISTERS, SPANISH ACADEMIC NETWORK.

Figure 4.6 shows the phase-plane plot for a set of throughput registers corresponding to 5 days of traffic in the Spanish academic network. This plot is numerically obtained, using a first-order finite difference method to approximate the derivative function. Note that, apart from the points with abnormal throughput values, there are observations that present velocities (*i.e.* throughput variations) far from the typical joint values.

Meanwhile, Figure 4.7 shows the phase-plane plot for the same set of throughput registers obtained with analytical differentiation applied on the functional representation using third grade splines. Notice that some points are not considered in the interpolation of the curves, eliminating in this case the abnormal points observed in the previous plot.

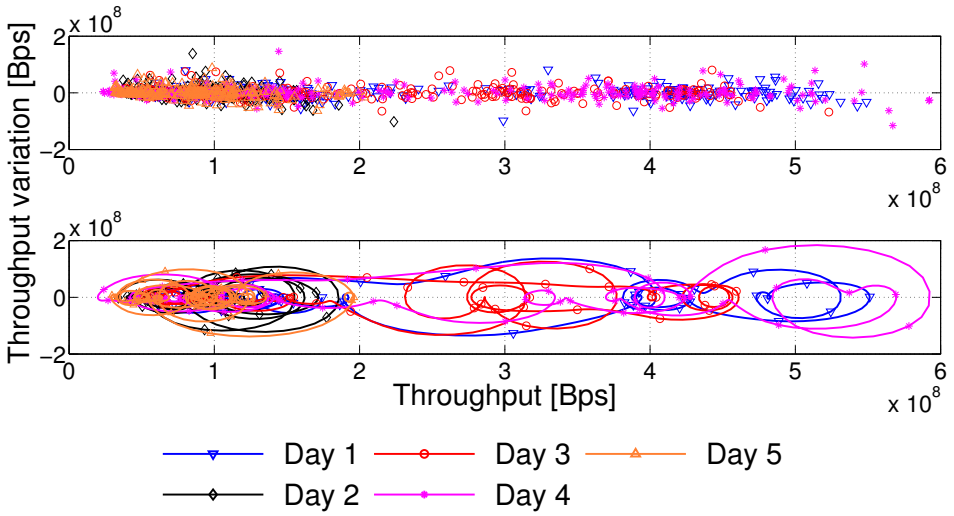

FIGURE 4.7: COMPARATIVE VIEW OF PHASE-PLANE PLOT OBTAINED WITH POINT ESTIMATION (FINITE DIFFERENCE METHOD), AND ANALYTICAL DERIVATION BASED ON B-SPLINE REPRESENTATION.

This representation is useful for visual detection of common events and provides extended information about the evolution of the network state. Additionally, the inclusion of several parameters in this analysis, which is straightforward given the functional definition of curves and surfaces, allows the study of joint relations between different magnitudes and their derivatives.

Given that FDA includes clustering and classification methods for curves and surfaces, phase-plane representation can be used to characterize different events in network dynamics. The study of behavioral aspects of network management data represented in such a form is, as a result, an interesting field of study that could produce solutions that overcome limitations derived from anonymity issues and data encryption.

Additionally, with the flexible network infrastructures such as cloud deployments, dynamical study of network behavior can solve different matters. For example, in [62] authors provide a recent application of dynamical analysis based on neural networks (which can be used to approximate functions) and catastrophe theory to detect anomalies in cloud environments. With the current interest in this type of deployments, there is room for further research in this area with a purely functional approach, as it could provide enhanced results.

4.5 Functional depth for network management data

Functional depth analysis provides a low computational cost alternative to obtain some order statistics between the curves/surfaces that represent parameters/set of parameters. The definition of these order statistics gives robustness when analyzing the typical behavior of a network, as a result of the isolate character of outliers and abnormal values.

In figures 4.8 and 4.9, we show the results of depth analysis on the time series. In Figure 4.8, a depth region based on the definition given in Equation 2.22, covering the 80% of observations of an extended set of throughput registers from the Spanish academic network. The lines marked with the black triangles correspond to the curves that delimit that region. The mean curve is represented in red with asterisks, while the blue lines with circles indicate the confidence interval of the mean at each point with $\alpha = 5\%$. The depth region leave inside this values, providing a frontier based on the structure of the functional curves that represent the data. In Figure 4.9, which includes a very similar summary, the most remarkable fact is the robustness of the depth-based results.

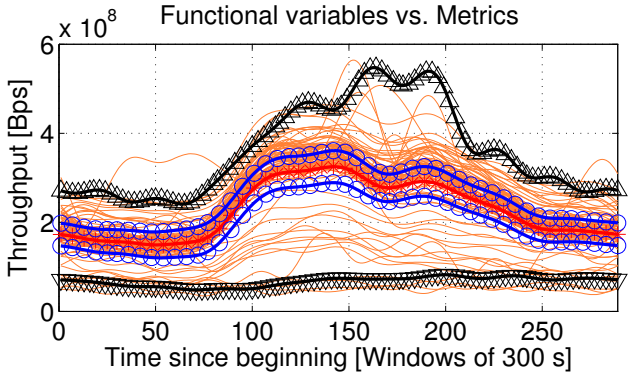

FIGURE 4.8: EXAMPLE OF DEPTH REGION USING AN EXTENDED SET OF THROUGHPUT REGISTERS, ACADEMIC NETWORK. MEAN CURVE: RED ASTERISKS. MEAN CONFIDENCE INTERVAL: BLUE CIRCLES. DEPTH REGION: BLACK TRIANGLES.

The appearance of network infrastructures that allow both dynamic configuration of rules and resource deployment (*e.g.* SDN (Software Defined Networking) or the ABNO (Application-based Network Operations) architecture [63]) points to the establishment of baselines that take into account network behavior at each time frame. Depth-based metrics are good candidates to the definition of such baselines, as they allow the construction of regions covering a certain proportion of the observed curves. Furthermore, the multi-

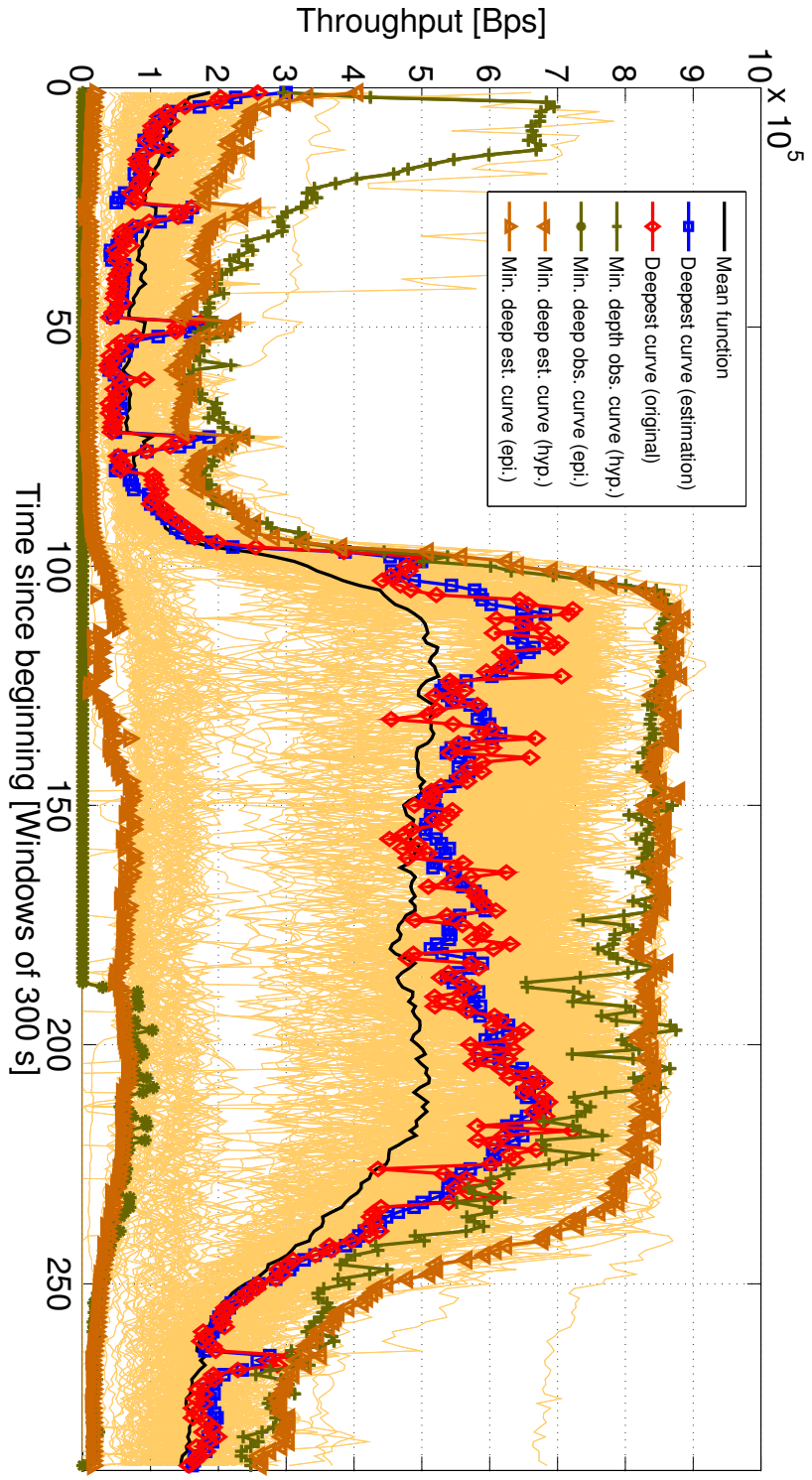

FIGURE 4.9: SUMMARY OF DEPTH ANALYSIS FOR THE SECOND DATASET.

variate definitions support the joint consideration of a set of parameters, which is interesting as some events required the monitoring of several characteristics to be detected [47].

4.6 Functional homogeneity for network management data

The use of functional homogeneity statistics in the study of network management data is a natural approach to contrast the hypothesis of the representativeness of a certain set of parameters. Thus, using statistics as the presented in [27], we can define an automatic algorithm to detect those parameters that characterize the typical state of a network — those that have a homogeneous behavior in periods.

In network management activities, functional homogeneity tests can be applied in order to detect changes with different aims. These tests allow the definition of new metrics based on invariant parameters taking into account their whole evolution and not only statistical summaries – *e.g.* means or medians. Furthermore, functional homogeneity can help to detect anomalous events, as they produce divergences in observations that can be undetected in aggregated metrics. Finally, sustained trend changes can be easily detected with the analysis of the separation between curves of evolution.

4.7 Other activities

As mentioned in Chapter 2, there are other topics in FDA that can be applied to network management activities. For example, the application of clustering and classification of curves to the identification of certain characteristics of traffic (*e.g.* application that generates that traffic) must be studied, in order to overcome limitations derived from data encryption. Forecasting based on functional regression is, additionally, of particular interest, as it could impact on dynamical planning in SDNs and other flexible network and systems, such as in Cloud infrastructures.

4.8 Conclusions

In this chapter, we have shown the results of applying FDA to several processes of the NTMA scope. In this case, we have focused on the analysis of network measurement time series. Our results have shown that FDA provides promising results when considered in several network management activities. First of all, functional representation opens the gate to a novel consideration of compression and warehousing for network measurement

time series. The use of FPCA enables to further reduce storage necessities for such data, and provides means to cluster, classify and detect anomalous events from network measurements. Phase-plane analysis is also interesting to study and represent the state of network deployments, particularly if we consider flexible infrastructures. Other techniques with obvious applicability in network management activities, are out of this first study. We plan to further study them as future work.

5 Network flow data and FDA

5.1 Introduction

In this chapter, we propose a simpler and more flexible way to summarize the network flow behavior. This approach improves the definition and visualization of network flow categories defined in terms of a statistical analysis of their characteristics. To define such categories, we consider certain probability levels in a CDF estimation of network flow characteristics. Then, those probability levels are mapped to a set of flow characteristic values via the probability integral transform. To obtain the CDF estimation, we rely on a FDA-based approach using ECDF observations obtained from network flow records.

First, we motivate the method, stating the objectives behind it. Then we present the fundamentals of our proposal, with a further description of the problem we are facing. After that, we provide a review of the probability results which are the ground of our solutions and explain the different FDA steps we apply to obtain a representative CDF estimation. Finally, we provide a empirical evaluation of such method and extract some conclusions from this chapter.

5.2 Motivation of the method

Our proposal is to conform categories for different flow characteristics in terms of different probability levels in the CDF via the probability integral transform. This approach entails different advantages. On the one hand, the use of characteristics at flow-level improves the analysis that can be done if we use more aggregated data, and does not incur in privacy issues. On the other hand, our approach induces a methodology to study and understand the flows traversing the network. Specifically, with the use of these statistical summaries we open the door to the elaboration of *(i)* tests to detect changes and events by dealing with all the variables under study in the same manner, and *(ii)* a novel representation to present the evolution of network state to managers.

This can be useful to visualize the traffic evolution, and easily detect changes in its pattern. Bearing in mind that losses in the summarized information can lead to restricted or even erroneous conclusions, our approach solves this problem by defining a set of intervals related to certain probability levels using the probability integral transform [64]. Such transformation ensures that given a set of samples of a concrete characteristic, the distribution on equally-spaced ranges (*i.e.*, quantiles) over the cumulative distribution function of the sample will be uniformly distributed. Intuitively, this means that if we take, for example, the empirical percentiles of a sample and we count the values appearing between each percentile, we will approximately obtain the same figures.

With respect to common traffic throughput time series, the representation over time of these set of values would provide a richer view of the network traffic, which is at the same time easy to understand by a network manager. If a change on the behavior occurs, it would break the uniform distribution over the intervals vector and a change will be detected. Particularly, we claim that the detection of excursions on uniformly distributed values is easier than other approaches for both automatic tools and network managers. First, it is trivial to use a contrast hypothesis test for uniformity (*e.g.*, χ^2 test) —however, as we will explain, some limitations apply for non-continuous samples. Second, network managers can also easily detect excursions from uniformity after a simple visual inspection. Note that each defined interval (hereafter, a category) must fairly show the same number of samples and if we plot that over time, the results will be represented as equispaced curves. Otherwise, the uniform distribution is not being fulfilled. This observation gave rise to the framework `Dictyogram`, which allows network managers to visually inspect the output of our approach.

5.3 Applying FDA to the study of network flow characteristics

5.3.1 Formal description of the method

Our goal is to describe flow characteristics in terms of a summarized representation using the CDF. To do so, we define categories using the probability integral transform [64]:

Theorem 2. Probability integral transform: *Let X be a continuous random variable with cumulative distribution function F_X. Then $F_X(X)$ follows a uniform distribution on $[0, 1]$.*

Therefore, to obtain the summary, we consider the distribution of values in $F_X(X)$ and select a certain partition of data defined by a set of probability levels $\{P_i\}_{i=1...n}$. Hence, the flow categorization is given in terms of a corresponding set of values $\{C_i\}_{i=1...n}$, which are defined as

$$C_i = F_X^{-1}(P_i) \tag{5.1}$$

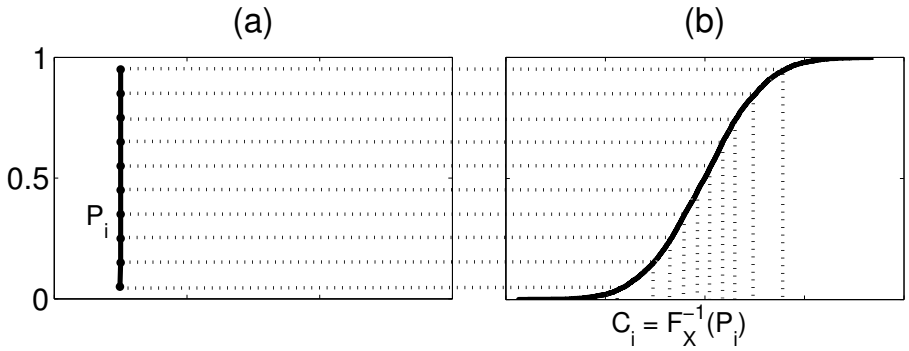

FIGURE 5.1: DEFINITION OF CATEGORIES IN TERMS OF A SET $\{P_i\}_{i=1...n}$ OF PROBABILITY VALUES, WITH THE CORRESPONDING CATEGORICAL DATA $\{C_i\}_{i=1...n}$ WITH $C_i = F_X^{-1}(P_i)$.

Needless to say, the width between the C_i corresponding to each quantile is not equal, as those part of the random variable with little probability mass will define large intervals, thus compensating those part of the variable with large probability mass. Then, the set of values that makes up the vector of intervals will define a signature of the behavior of a given characteristic.

In Figure 5.1 we illustrate the meaning of Equation (5.1). We link the category frequency behavior with the value holding this accumulated probability via the cumulative distribution function F_X. For instance, Figure 5.2 shows the application of this theorem using 5000 realizations of a random variable following a normal distribution with parameters $\mu = 30, \sigma = 1$. In this figure, we represent in (a) a histogram of 10 bins of the values of $F_X(X)$, and in (b) the ECDF of the sample. Additionally, we have tested that a different numbers of bins does not induce changes in the behavior of the histogram of $F_X(X)$, which remains uniform. Given that the hypothesis of continuity of the theorem is met, the result holds in this case.

It is clear that we can use the quantiles of network flow characteristics to define categorizations with a uniform distribution of flows for each category. Additionally, as a result of the definition of quantile, we can state that if the number of network flows is stable, then these two situations will be equivalent:

- A change in the number of network flows in the category whose extreme values are defined by two given quantiles.
- A change in the values of those quantiles.

There are several advantages derived from the definition of these network flow categories. As the distribution of the number of flows for each category is uniform, it is easier to

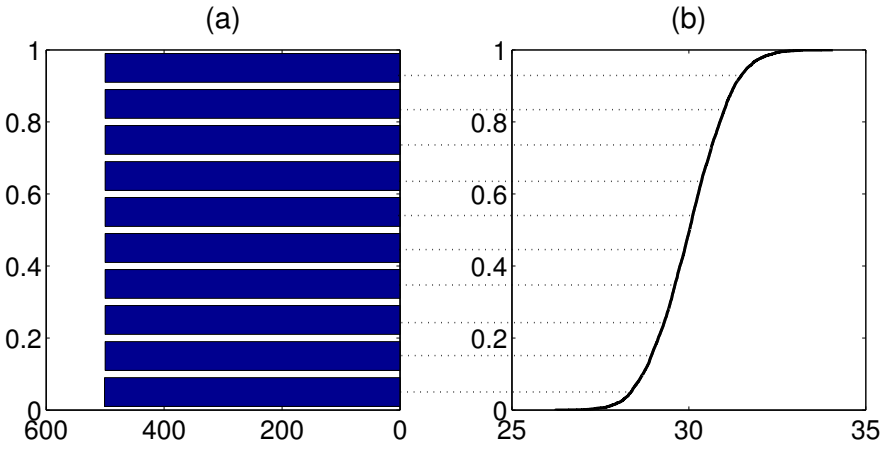

FIGURE 5.2: BEHAVIOR OF (A) THE HISTOGRAM OF $F_X(X)$ AND (B) THE ECDF OF X FOR 5000 REALIZATIONS OF A NORMAL RANDOM VARIABLE OF PARAMETERS $\mu = 30, \sigma = 1$.

represent the behavior of network flows. Moreover, it is possible to detect changes using a homogeneity test (*e.g.*, Pearson's χ^2 test). Additionally, our approach provides notions about the position of each category inside of the set of observations, which is interesting as usage patterns are related to the characteristics of flows [39].

Nevertheless, three main issues arise during the practical application of this method in network studies, which are later solved:

1.– *Human network managers* can barely cope with the joint analysis of a large number of categories. This fact makes necessary the definition of representative summaries that allow the interpretation of network measurement data.

2.– It is not usual to know the cumulative distribution function of empirical observed random variables, so it is necessary to estimate such functions.

3.– The continuity of random variables hypothesis is not always met by network flow parameters (*e.g.* flow size in bytes is an integer value). If we are using characteristics which are not continuous random variables, the uniformity of quantiles could not be hold.

With respect to this last issue, the definition of a uniformly distributed categorization of network flows can be really challenging if the measurement process includes any sampling (*e.g.* packet sampling), as we will show in Section 5.5. To illustrate the absence of uniformity in the values of $F_X(X)$ if X is not continuous, in Figure 5.3 we show the behavior of 5000 realizations of a random variable following a Poisson distribution with parameter $\lambda = 30$. The meaning of each subplot is equivalent to those in Figure 5.2. Note

that, if the distribution is discrete, the mass distribution of X is very concentrated, thus, the histogram of $F_X(X)$ shows a small number of values for each bin.

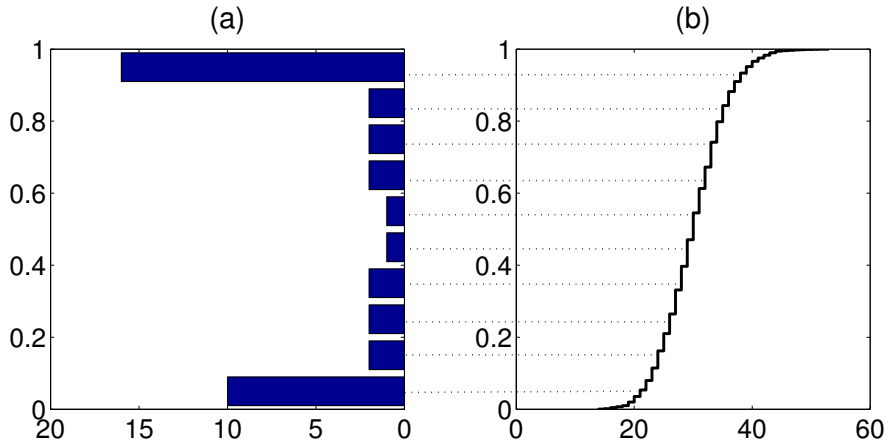

FIGURE 5.3: BEHAVIOR OF (A) THE HISTOGRAM OF $F_X(X)$ AND (B) THE ECDF OF X FOR 5000 REALIZATIONS OF A NORMAL A POISSON DISTRIBUTION WITH PARAMETER $\lambda = 30$.

5.4 Dictyogram

In this section, we will describe the characteristics of `Dictyogram`, our novel framework for the analysis and visualization of flow characteristics. `Dictyogram` is based on previously described method, and it is conceived to provide a detailed representation of the network state in an "manager-friendly" fashion. We have chosen this name because we aim at obtaining graphical results that can be like a network ($\delta\acute{\iota}\kappa\tau\upsilon o$ in Greek) electrogram, showing its vital signs.

5.4.1 Dimensionality reduction

Taking into account the limitations that we have explained for a direct application of the probability integral transform, we leverage some dimensionality reduction techniques to overcome these potential matters.

First, we regard to non-continuous random variables issue. In this scenario, the selection of $\{P_i\}_{i=1...n}$ can be tuned to minimize the impact of discontinuities of the CDF. It is important to note that those discontinuities are caused by values of X having large

probability mass. As a result, the maximum mass of a point restrict the cardinality of the set $\{P_i\}_{i=1...n}$ for a categorization that distributes uniformly the number of flows between categories. If we denote the maximum mass of a point as F_0, then n is bounded by $1/F_0$. Thus, taking into account the CDF estimation and this restriction, it would be possible to select a categorization holding the maximum resolution achievable.

Nevertheless, sometimes it will not be possible to define any categorization having this property —*e.g.*, think of a random variable taking a value with a probability greater than 0.5. Still, in this worst case scenario our proposal to define categories can be useful, even without achieving a strict uniform distribution of flows. The summarization of network behavior, and the visualization and study of network dynamics are interestingly enriched, as we will show in Section 5.5.

Regarding the results presented to network managers, the dimensionality reduction that `Dictyogram` provides entails other advantages. One of the definitions of visualization states that it is *"a cognitive process performed by humans in forming a mental image of a domain space"* [65]. Thus, the data obtained after applying the probability integral transform must be presented to users in such a way that they can comprehend the characteristics of the system under analysis. `Dictyogram` lets control the resolution of the visualization of the distribution that network flow characteristics follow. Additionally, if we obtain time series representing the number of flows in each category, we will have temporal snapshots of the distribution evolution of the characteristic under analysis. Moreover, other high-dimensional visualizations can be obtained using suitable graphical representations, for instance, heat maps.

5.4.2 Estimation of the cumulative distribution function

Although the Glivenko-Cantelli theorem [66] assures that the empirical estimation of the ECDF converges to the CDF as the number of observations increases, our goal here is to use the ECDFs observed in different days without accumulating all the values of the characteristic under analysis. That is, we map the analysis of a certain flow characteristic into the FDA setup via the estimation of its CDF. This methodology is more scalable when considering long-term studies, as the amount of required data is drastically reduced —*e.g.*, we can only keep a certain number m of points for each ECDF, instead of all the observations for each flow characteristic value.

To estimate the cumulative distribution function of the flow characteristic under analysis, we discuss three different approaches, namely *(i)* to use the mean function of the observations, *(ii)* the deepest observation, or *(iii)* the curve that maximizes the functional depth. Let us describe each of these approaches and highlight their main advantages and shortcomings.

First of all, we consider the use of the mean function of observations. That is, given a

set of observations of the ECDF of the characteristic under analysis, which we represent as $\{F_{X_i}\}_{i\in 1\ldots n}$, we define our model as

$$F_X^{mean} = \frac{1}{n}\sum_{i=1}^{n} F_{X_i} \tag{5.2}$$

Given that all elements in $\{F_{X_i}\}_{i\in 1\ldots n}$ are well defined, so it is F_X^{mean}. This approach provides a solution with reduced computational cost, which can be valuable in some scenarios. Nevertheless, the use of the mean as a central tendency measure is not a robust approach. As a result, if there are outliers or heterogeneous behaviors in $\{F_{X_i}\}_{i\in 1\ldots n}$ (*e.g.*, different distributions between weekdays and weekends), the model would be deviated and bad-representing the distribution function, as we will illustrate in Section 5.5. Moreover, problems with integer values for certain flow characteristics arise when using this approach. In fact, it is difficult to describe how to manage rational values in this context, and it can lead to incorrect behaviors of the model.

To cope with these matters, we describe now two alternatives that provide a more robust approach and avoid problems with values out of the domain of definition of the observations. Our proposals are defined in terms of *functional depth*.

Using the definition of functional depth that we have introduced in Equation 2.22, our second alternative is to obtain the deepest observation of the sample set. We can find the observed ECDF with the highest value of $MS_{n,H}(x)$. This approach entails to increase the computational cost of the estimation, but at the same time it is protected against outliers in global terms; this is, ECDFs that are far from the usual observed behavior (for instance, ECDFs from atypical days).

A third alternative arises by following the notion of centrality. We see that the median curve (that is, the curve that passes through the median value at each probability level) is the function that maximizes Equation (2.22) when considered at each point. As a consequence, this approach captures the typical behavior of the flow characteristics at each probability level. This is the most computationally demanding approach that we are considering. It provides a robust exploration of the typical behavior locally (instead of globally as in the case of the deepest observation) but it needs to order all the observations at each probability level.

The next section empirically evaluates these three approaches. Our findings point to assess the alternatives for each particular deployment scenario, as the properties of traffic characteristics may induce changes to the behavior of ECDF estimations.

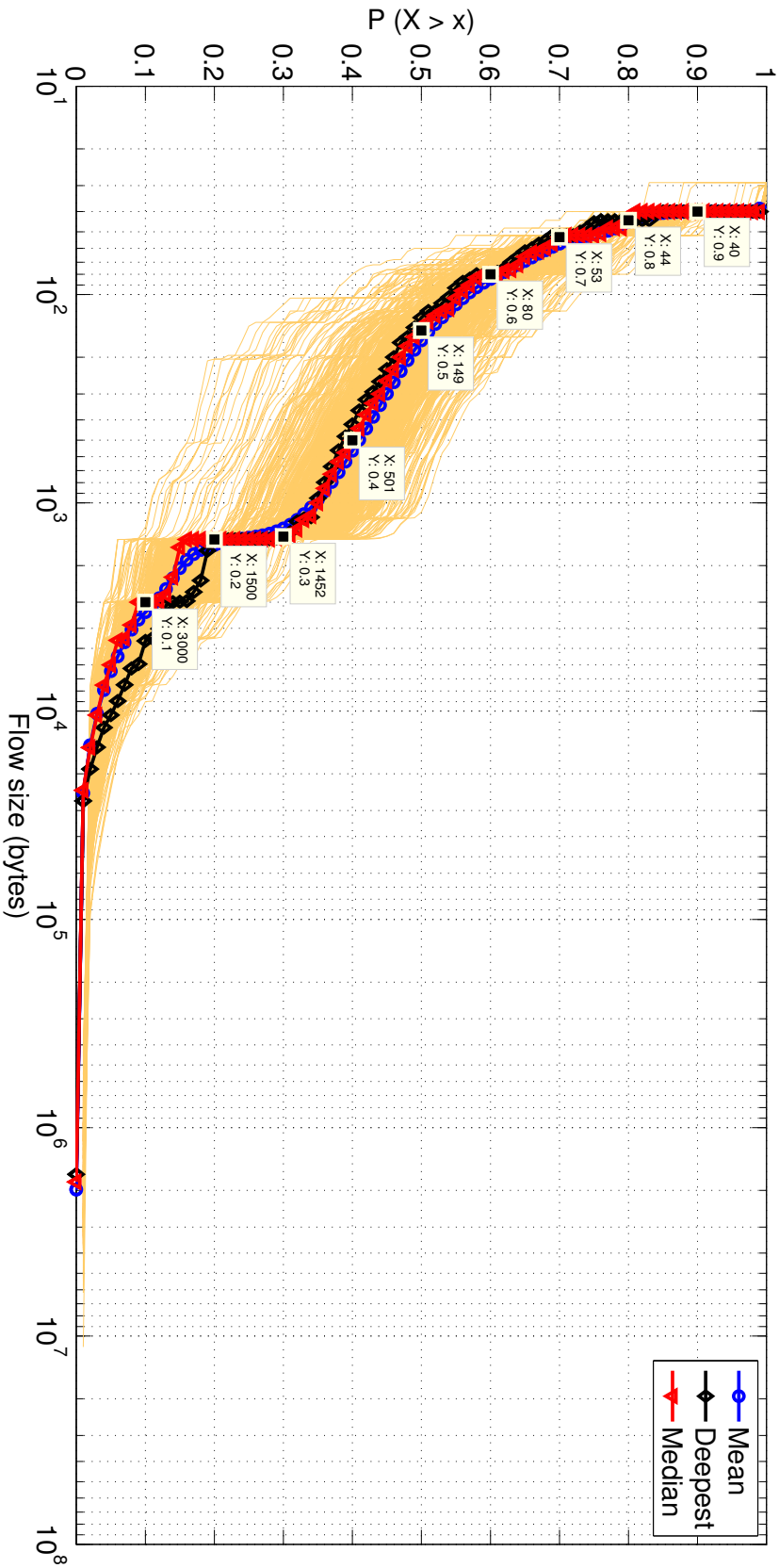

FIGURE 5.4: COMPARISON BETWEEN OBSERVED CCDF (ORANGE LINE, NO MARKER) FOR EXPORTER A, AND MODELS OBTAINED USING THE MEAN (BLUE LINE, CIRCLES), DEEPEST (BLACK LINE, DIAMONDS) AND MEDIAN (RED LINE, TRIANGLES) FUNCTIONS.

5.5 Case study

In this section we will present a case study in which we put into practice the ideas behind this article. To lead this study, we have used `Dictyogram` to process network flows from 5 different flow exporters of the Spanish Academic Network during a period of four years, from 2007 to 2011. Some other results and information can be found in [67]. Given that we will study the underlying distribution function of these exporters, it is important to note that traffic has been sampled at a rate of one out of 100 packets. In our case study, we have decided to present an example studying network flow size in bytes, but the techniques presented here can be used for any other network characteristic, such as packets or duration of the flows.

For each of the exporters (A, B, C, D and E hereafter), we have calculated the deciles of the flow size in bytes using the three methods described in the previous section, obtaining the results shown in Table 5.1. We can appreciate the effect sampling exerts on the underlying data, as usually the first three deciles are in the 40-60 bytes range. Further analysis of the data also shows that up to 90% of the sampled flows per day consist of only one packet, following the clue given by the 9th decile which is always close to 1500 bytes —one TCP packet with full payload. This means that barely 10% of the sampled flows per day have two packets, and due to the nature of the Internet traffic, we find high variances in the middle deciles.

Sharing the conclusions presented in [68], we have also assessed that the exporters do not share values for the deciles (although all of them presented similar intrinsic features), confirming that measurements collected in a network could not be extrapolated to others. It is recommended to use at least one month worth of data in order to obtain the deciles. These measurements should be recalibrated every so often, because traffic behavior may alter them significantly enough.

Additionally, we have plotted in Figure 5.4 the results for exporter A, showing the model obtained with each method. We have selected this exporter because of the high variability of the daily ECDFs, which leads to noticeable differences in the derived models —note that axis of abscissas is in logarithmic scale. In this figure, we have also labeled the decile values in the estimation that presented the best behavior for this exporter after the empirical evaluation presented below.

To measure which of the three methods distributes more uniformly the flows, we have calculated the Pearson's test-statistic for all three methods during a period of 4 weeks in every exporter. The results are depicted in Figure 5.5, and summarized in the last column of Table 5.1.

As we can see, no method outperforms the other two, as the results change among the observed exporters. Depending on the level of aggregation of the exporter under study one might prefer one or the other, bearing in mind that the median is the most computa-

TABLE 5.1: DECILES OBTAINED USING THE ESTIMATION OF THE CUMULATIVE DISTRIBUTION FUNCTION WITH EACH APPROACH.

Exporter	Method	Deciles (bytes)									# Best*
A	Mean function	40.019	44.88	57.047	84.18	165.99	562.13	1327.8	1595.6	3348.8	0
	Deepest obs.	40	44	52	80	129	420	1448	1500	4600	3
	Median function	40	44	53	80	149	501	1452	1500	3000	25
B	Mean function	39.982	47.244	59.644	93.771	211.99	824.68	1467.5	1582.3	3794.3	0
	Deepest obs.	40	52	64	92	163	531	1420	1500	4476	6
	Median function	40	48	60	92	208	833	1480	1500	3744	22
C	Mean function	39.817	45.583	51.782	72.296	124.01	346.59	1148.5	1486.6	3028.3	20
	Deepest obs.	40	48	52	70	120	312	1152	1500	3000	8
	Median function	40	46	52	74	122	348.5	1260	1500	3000	0
D	Mean function	39.914	43.36	53.505	82.337	165.01	485.46	1329.9	1508.4	3991.9	0
	Deepest obs.	40	49	60	86	146	355	1420	1500	3604	23
	Median function	40	44	52	80	160	496	1420	1500	4170	5
E	Mean function	40	46.415	62.596	95.141	180.35	654.24	1404.5	2117.3	4736.7	0
	Deepest obs.	40	51	63	93	160	367	1260	1840	5680	28
	Median function	40	48	62	91	168	600	1420	2120	4260	0

* **# Best** column shows the number of days in 4 weeks that each method provided the best Pearson's test-statistic value.

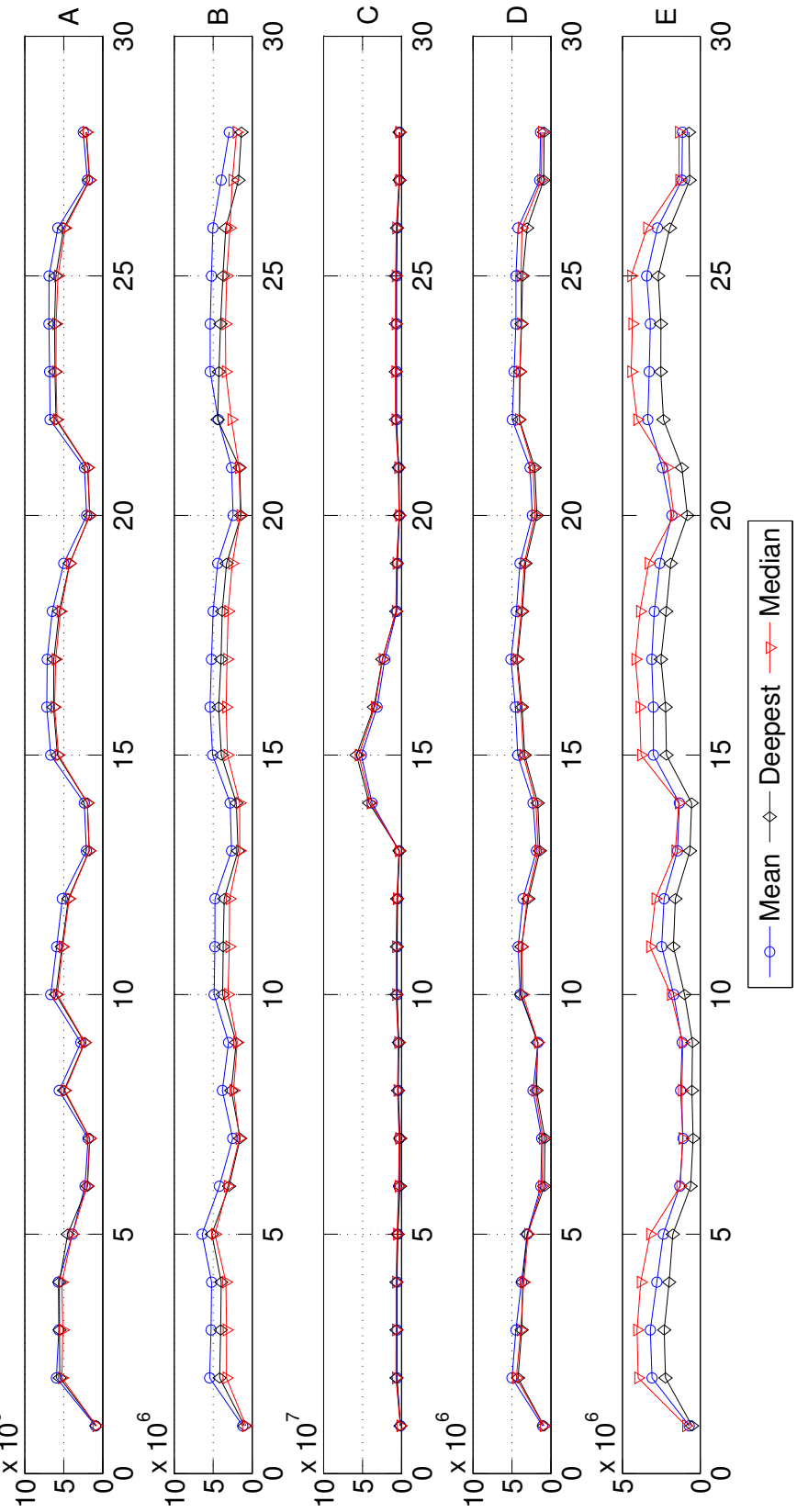

FIGURE 5.5: EVOLUTION OF THE PEARSON'S TEST-STATISTIC FOR ALL EXPORTERS DURING OCTOBER 2010. (LESS IS BETTER.)

tionally expensive of the three methods presented. In exporters with less aggregation (and thus more variance) such as E, the median yields better results. Exporter A also presents less aggregation, but the deepest observation in this case is consistently the method with lowest Pearson's test-statistic. Looking at the evolution of the Pearson's test-statistic for Exporter A, we can see that the deepest observation test-statistic value is fairly similar to that of the median, and it could be argued that in general it would seem more reasonable to use the deepest observation, as the median is the most computationally expensive method. In cases with low aggregation one should not use the mean to obtain the deciles, due to higher variance. In cases where there is more aggregation (exporters B, C and D) we have a similar situation. We can safely say that, although the mean is the cheapest method it does not yield the best results for uniformity overall, and that the median and the deepest observation do not seem to differ significantly. Nonetheless, the results obtained point out that one or other method should be considered depending on the aggregation of the exporter.

The goal of the visualization produced by `Dictyogram` is to present the number of flows between each interval defined by the deciles. If $\{d_k\}_{k=1...9}$ are the deciles obtained through a given method, then we define the intervals as $[0, d_1] \cup (d_j, d_{j+1}] \cup (d_9, \infty)$ for $j = 1 \dots 8$. For each of these intervals we will present a plot $f_i(t)$ for $i = 1 \dots 10$ that will represent the number of active flows whose size is within its given size interval at a given time t. In our study, we have chosen visualizations of one day and granularity of one minute.

As stated in Section 5.3, it is not trivial to obtain a uniformly distributed categorization of the flows because of the sampling. As *mice* flows tend to be more present than *elephant* flows, smaller flow size categories have an inherently higher number of flows than larger ones. Furthermore, flows from determined sizes are more likely to appear than others (40, 48, 1500, etc.). A categorization defined with the deciles, as explained, further impedes uniformity of distribution among categories, as usually the deciles concur with these sizes. Nevertheless, it does not impede a good visualization.

We present a `Dictyogram` representation example in Figure 5.6. The represented data corresponds to one day (26/12/2011) and one particular exporter (B). The values of the deciles for such exporter B are those presented before in Table 5.1 for each method, where median and deepest observation provide the best flow classification. As shown, the use of the mean does not work correctly with smallest deciles, which become overlapped. In this figure, we have plotted the functions $f_i(t)$. To improve visualization, we stacked each $f_i(t)$ function, so that lowest size interval is plotted at the bottom and so on. The accumulation of the $f_i(t)$ functions provides several advantages. It provides a clear understanding of what is happening in the network at any given time. With a quick glance one can understand how the traffic is distributed, and which size intervals are responsible for the majority of traffic observed at any given time.

Importantly, it can be used as a tool to detect anomalies in the network. This day presents a large anomaly from 17:00 to 19:00 approximately, that is, an unusually high

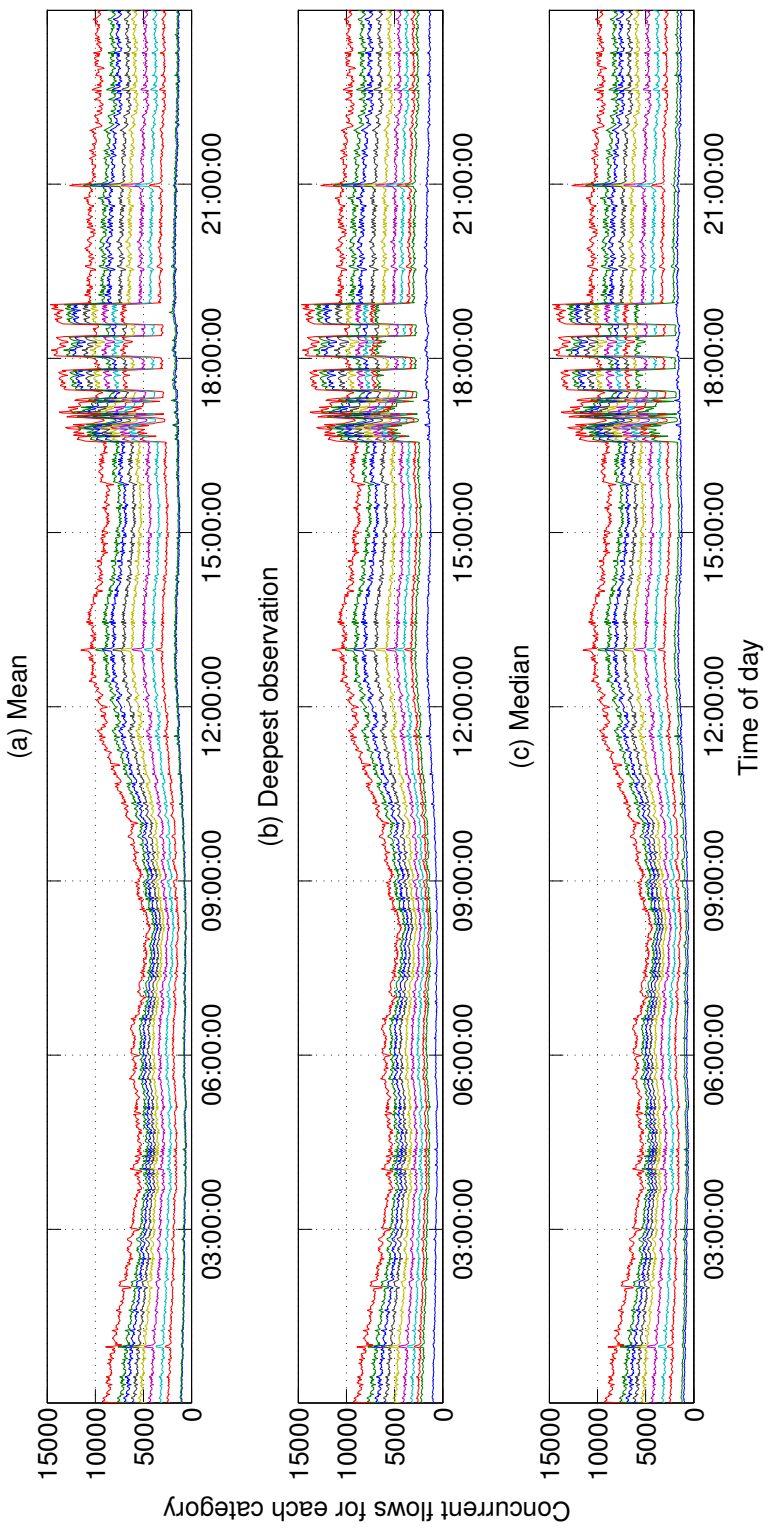

FIGURE 5.6: DICTYOGRAM REPRESENTATION OF $f_i(t)$ WITH THEIR RESPECTIVE SIZE INTERVALS DELIMITED BY THE DECILES GIVEN BY *(a)* MEAN, *(b)* DEEPEST OBSERVED ECDF, AND *(c)* MEDIAN.

number of active flows during that time. Thanks to the accumulated plots we can easily identify the size interval (or intervals) that are causing a spike at any given time. In this case, flows between 41 and 52 bytes. Looking at $f_{10}(t)$, which encloses all cases, it is possible to spot this anomaly, and the categorization let us find which flows are responsible for the alteration in traffic. Given the shape of the anomaly and the size interval of the anomalous flows an analyst can rapidly build up an idea of what might be happening. Following our example, one might hypothesize that we are watching some kind of port scanning or SYN flooding attack. One fast query to the flow records confirms that between 17:00 to 19:00 we were witnessing a port scanning attack, along with a massive attack via SSH (283453 SSH flows in two hours) from one specific host in China. The flow size was 48 bytes, corresponding to a SSH start-up session message. Additionally, other less obvious anomalies (when looking just at $f_{10}(t)$) occur that day, such as those happening around 1:00, 13:00 and 21:00, which can be easily identified with the `Dictyogram` representation.

All anomalies shown in Figure 5.6 can be automatically detected using time series filters, such as a Holt-Winters filter [69] or other exponential smoothing algorithms. These algorithms are not new for the network community, as the Jacobson algorithm for TCP round-trip time estimation [70] follows a similar approach. Smoothing each $f_i(t)$ and setting confidence intervals will pinpoint the anomalies, making the less obvious ones visible to analysts.

5.6 Conclusions

In this chapter, we have proposed a novel method to summarize, analyze and represent network flow characteristics. This method provides several advantages in various network management activities. It eases network data visualization, anomaly detection, and provides a methodology to define signs and network flow categories with semantic criteria based on the statistical properties of observations. To derive such method, we have used the integral probability transform and a FDA-based estimation approach, which shows the advantages derived of the application of this latter analytical framework in the NTMA domain. We have also presented a tool implementing this technique, `Dictyogram`, which has been used to empirically evaluate our solution in an extensive case study. Making use of this system, we have shown the usefulness of our method to obtain a network behavior characterization and to detect anomalous events in a real operational network.

6

CONCLUSIONS

6.1 Summary

We have presented an evaluation of the advantages that FDA entails in the area of Network Management. Our study has provided an initial exploration of this set of statistical techniques that motivates a deeper insight to its applicability and further evaluation of the results that can be achieved when using these techniques. We have followed two lines, which are related to the study of network measurement time series and network flow characteristics, respectively.

Regarding to the study of network measurement time series, our analysis has focused separately on different FDA aspects, namely data representation, FPCA, phase-plane analysis, functional depth and homogeneity. We have offered a formal description of several elements involved in these topics, providing a mathematical motivation for the advantages and soundness of our proposal.

We have linked these statistical concepts to certain problems that arise in many Network Management activities, analyzing network measurements time series. We believe that this initial exploration of the applicability of FDA to network management tasks indicates the direction of an interesting step forward for researchers and practitioners in the area of Network Management and analysis.

Regarding to the study of network flow data in terms of ECDFs, we have devised a method that help in network management tasks. Claiming that detection of changes in uniformly distributed values is more intuitive and giving to the visualization the importance it deserves, we have analyzed the different steps of the practical application of our method. We have explored its limitations, such as the problem of discretization and non-continuity of the random variables under analysis. Additionally, we have studied the estimation of the CDF of flow characteristics using observations of different ECDFs with a FDA-based approach. We have proposed three different approaches that obtain robust and representative results —namely, mean, deepest and median functions.

Finally, we have implemented our method in a framework, `Dictyogram`, available

under request. We have presented a real case study on flow data from the Spanish academic network to illustrate the usefulness of `Dictyogram`. Specifically, we have focused on flow sizes, but it is worth remarking that we could have studied any other network characteristic.

6.2 Contributions

We have reached several conclusions after the application of FDA to network measurements time series. Let us highlight some of them.

We have shown that functional representation, which is the initial point where applying functional analysis methods to empirically obtained data, offers new possibilities for data compression. This first-level compression is due to the reduction of the number of values that describe data when it is represented with respect to a functional basis.

The application of FPCA as a data preprocessing technique allows a second level of compression, as a result of the reduction of the size of the basis with a criterion based on the proportion of the original variance that is covered. Additionally, it provides a semantic decomposition of network parameters that enriches network dynamic interpretation.

Pointing to data visualization and analysis, we have considered phase-plane analysis as a functional instrument that provides a starting point for the development of solutions oriented to the characterization and detection of events and abnormal behaviors.

The notions of functional depth and homogeneity provide means to define baselines and to detect changes in network dynamics respectively. Both of them are robust approaches that consider curves or surfaces as a whole, allowing to take into account the joint behavior of an arbitrary number of parameters. As previously commented, this is of vital importance in certain situations that cannot be diagnosed without this multivariate vision.

Regarding our novel proposal to summarize network flow characteristics, we have evaluated its advantages and limitations during a real case study.

We have shown that the advantages of our method are manifold. First, it is more straightforward to apply than other approaches, as we use a simple vector to summarize the behavior of a network characteristic. Second, it allows the description of the temporal evolution of the flows traversing the network. Finally, the identification of changes on such a vector becomes trivial, as a simple visual interface lets network managers assess abnormal changes. Our case study has highlighted the applicability and ease of use of our approach.

Taking into account that the data we have used entails applying our method in the worst case scenario in terms of it limitations, we have shown that they do not significantly

hinder the results that can be obtained.

Additionally, we have applied three different approaches to estimate network flow size CDFs using ECDFs. Our discussion and results can be spread to other network flow characteristics. Moreover, this discussion could be of interest for other researchers, as it is possible to apply these novel statistical techniques to estimate other models.

To conclude, we have presented a set of tools and guidelines that can be applied during several analytical activities in the network management scope. Moreover, we have defined a novel data representation that can be successfully applied in many different network research tasks, being useful for both analysts and researchers.

6.3 Future work

This work contains the results of a first contact of FDA techniques with network management. It must be continued with an extended evaluation of results, with the development of advanced methods that take advantage of the principles exposed in this work, and further exploitation of the strengths of functional data analysis. Thus, our conclusions open several future work lines.

We plan to further evaluate compression capacities of the functional representation, and study relations between network anomalies and perturbations in the FPCA weights / coefficients. Additionally, we would like to carry out a purely functional study of network dynamics, which can be applied to detect security issues as some works point.

Another future study is related to how to summarize several different network behaviors in a multivariate uniform distribution. This is quite interesting, as it could span additional methods to detect changes and anomalies. We are also studying the possibility of modeling the categories presented in this work with other well-known distributions and not only as uniform signatures.

Additionally, regarding to the mining of network flow data, we plan to study the distribution of the Pearson's test-statistic to detect anomalous events. Moreover, we consider testing the stability of the estimation of the CDF, by defining some criteria to recalibrate the model. Another future work is the exploration of other representations with higher dimensionality —*e.g.*, heat maps, based for instance in percentiles.

A

APPENDIXES

A.1 Contributions

The results derived from this work are directly included in the following publications:

- **Chapter 4:**
  - **Functional Data Analysis: A step forward in Network Management.** D. Muelas, J.E. López de Vergara, J. R. Berrendero. Proceedings of the IFIP/IEEE International Symposium on Integrated Network Management, IM'2015, Ottawa, Canada, 11-15 May 2015.
  - **A novel statistical approach for the analysis of Network Monitoring time series.** D. Muelas, J. E. López de Vergara, J. R. Berrendero. Presented in 5th PhD School on Traffic Monitoring and Analysis, co-located with the 7th International Workshop on Traffic Monitoring and Analysis (TMA 2015), Barcelona, Spain, April 21-24, 2015.
  - **Análisis de Datos Funcionales para Gestión de Red: Técnicas, Retos y Oportunidades.** D. Muelas, J. E. López de Vergara, J. R. Berrendero, J. Aracil. Accepted for its publication in XII Jornadas de Ingeniería Telemática, JITEL 2015, Palma de Mallorca, Spain, 14-16 octubre de 2015.
- **Chapter 5:**
  - **Dictyogram: A statistical approach for the definition and visualization of network flow categories.** D. Muelas, M. Gordo, J.L. García-Dorado, J.E. López de Vergara. Submitted to 11th International Conference on Network and Service Management, CNSM'15, Barcelona, Spain, 9-13 November 2015.

Additionally, some contents of this work that have been commented in Section 3.3 are related to the following journal papers:

- **Multi-granular, multi-purpose and multi-Gb/s monitoring on off-the-shelf systems.** V. Moreno, P.M. Santiago del Rio, J. Ramos, D. Muelas, J.L. Garcia-Dorado, F.J.

Gomez-Arribas, J. Aracil International Journal of Network Management, 2014.

- **Low-cost and high-performance: VoIP monitoring and full-data retention at multi-Gb/s rates using commodity hardware.** J.L. García-Dorado, P.M. Santiago del Río, J. Ramos, D. Muelas, V. Moreno, J.E. López de Vergara, J. Aracil. International Journal of Network Management, 2014.

Bibliography

[1] M. V. Joseph, "Significance of data warehousing and data mining in business applications," *International Journal of Soft Computing and Engineering (IJSCE) ISSN*, pp. 2231–2307, 2013.

[2] S. Chakrabarti, M. Ester, U. Fayyad, J. Gehrke, J. Han, S. Morishita, G. Piatetsky-Shapiro, and W. Wang, "Data mining curriculum: A proposal (version 1.0)," *Intensive Working Group of ACM SIGKDD Curriculum Committee*, 2006.

[3] H. Shiravi, A. Shiravi, and A. Ghorbani, "A survey of visualization systems for network security," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 8, pp. 1313–1329, Aug 2012.

[4] L. H. Gibeli, G. D. Breda, R. S. Miani, B. B. Zarpelão, and L. de Souza Mendes, "Construction of baselines for VoIP traffic management on open MANs," *International Journal of Network Management*, vol. 23, no. 2, pp. 137–153, 2013. [Online]. Available: http://dx.doi.org/10.1002/nem.1820

[5] R. Hofstede, P. Celeda, B. Trammell, I. Drago, R. Sadre, A. Sperotto, and A. Pras, "Flow monitoring explained: From packet capture to data analysis with netflow and IPFIX," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 2037–2064, 2014.

[6] A. Bhardwaj and M. Singh, "Data mining-based integrated network traffic visualization framework for threat detection," *Neural Computing and Applications*, vol. 26, no. 1, pp. 117–130, 2015.

[7] E. Martinez Colomina, E. Fallon, M. Wang, and S. Fallon, "ADAMANT - An Anomaly Detection Algorithm for MAintenance and network troubleshooting," in *Proceedings of IFIP/IEEE International Symposium on Integrated Network Management*, ser. IM'2015, 2015.

[8] ITU-T Study Group 2, "TMN management functions," International Telecommunication Union, Recommendation M.3400, Feb. 2000.

[9] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, "Big data," *The management revolution. Harvard Bus Rev*, vol. 90, no. 10, pp. 61–67, 2012.

[10] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, Jan 2014.

[11] N. Aggarwal, A. Kumar, H. Khatter, and V. Aggarwal, “Analysis the effect of data mining techniques on database,” *Advances in Engineering Software*, vol. 47, no. 1, pp. 164 – 169, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0965997812000038

[12] J. Ramsay and B. Silverman, *Functional data analysis*, ser. Springer Series in Statistics. Springer-Verlag New York, 2005.

[13] A. Cuevas, “A partial overview of the theory of statistics with functional data,” *Journal of Statistical Planning and Inference*, vol. 147, no. 0, pp. 1 – 23, 2014.

[14] J. Ramsay, G. Hooker, and S. Graves, *Functional Data Analysis with R and MATLAB*. Springer New York, 2009.

[15] J. O. Ramsay, H. Wickham, S. Graves, and G. Hooker, *fda: Functional Data Analysis*, 2014, r package version 2.4.3. [Online]. Available: http://CRAN.R-project.org/package=fda

[16] P. H. Eilers and B. D. Marx, “Flexible smoothing with B-splines and penalties,” *Statistical science*, pp. 89–102, 1996.

[17] N. Locantore, J. Marron, D. Simpson, N. Tripoli, J. Zhang, K. Cohen, G. Boente, R. Fraiman, B. Brumback, C. Croux, J. Fan, A. Kneip, J. Marden, D. Peña, J. Prieto, J. Ramsay, M. Valderrama, A. Aguilera, N. Locantore, J. Marron, D. Simpson, N. Tripoli, J. Zhang, and K. Cohen, “Robust principal component analysis for functional data,” *Test*, vol. 8, no. 1, pp. 1–73, 1999. [Online]. Available: http://dx.doi.org/10.1007/BF02595862

[18] D. Bosq, *Linear Processes in Function Spaces. Theory and Applications.*, ser. Lecture Notes in Statistics. Springer Berlin, 2000, vol. 149.

[19] T. Hastie and W. Stuetzle, “Principal curves,” *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.

[20] P. Delicado, “Another look at principal curves and surfaces,” *Journal of Multivariate Analysis*, vol. 77, no. 1, pp. 84–116, 2001.

[21] U. Ozertem and D. Erdogmus, “Locally defined principal curves and surfaces,” *The Journal of Machine Learning Research*, vol. 12, pp. 1249–1286, 2011.

[22] K. Mosler, “Depth statistics,” in *Robustness and Complex Data Structures*, C. Becker, R. Fried, and S. Kuhnt, Eds. Springer Berlin Heidelberg, 2013, pp. 17–34. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-35494-6_2

[23] S. López-Pintado and J. Romo, “On the concept of depth for functional data,” *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 718–734, 2009.

[24] S. López-Pintado and J. Romo, "A half-region depth for functional data," *Comput. Stat. Data Anal.*, vol. 55, no. 4, pp. 1679–1695, Apr. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.csda.2010.10.024

[25] G. Claeskens, M. Hubert, L. Slaets, and K. Vakili, "Multivariate functional halfspace depth," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 411–423, 2014.

[26] ——, "Multivariate functional halfspace depth," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 411–423, 2014.

[27] R. J. F. Díaz, R. E. Lillo, and J. Romo, "Homogeneity test for functional data based on depth measures," Tech. Rep., 2014.

[28] P. Delicado, "Functional k-sample problem when data are density functions," *Computational Statistics*, vol. 22, no. 3, pp. 391–410, 2007. [Online]. Available: http://dx.doi.org/10.1007/s00180-007-0047-y

[29] D. Pigoli and L. M. Sangalli, "Wavelets in functional data analysis: estimation of multidimensional curves and their derivatives," *Computational Statistics & Data Analysis*, vol. 56, no. 6, pp. 1482–1498, 2012.

[30] A. M. Alonso, D. Casado, and J. Romo, "Supervised classification for functional data: A weighted distance approach," *Computational Statistics & Data Analysis*, vol. 56, no. 7, pp. 2334–2346, 2012.

[31] A. Cuevas, M. Febrero, and R. Fraiman, "Robust estimation and classification for functional data via projection-based depth notions," *Computational Statistics*, vol. 22, no. 3, pp. 481–496, 2007. [Online]. Available: http://dx.doi.org/10.1007/s00180-007-0053-0

[32] L. A. Garcia-Escudero and A. Gordaliza, "A proposal for robust curve clustering," *Journal of Classification*, vol. 22, no. 2, pp. 185–201, 2005. [Online]. Available: http://dx.doi.org/10.1007/s00357-005-0013-8

[33] L. Slaets, G. Claeskens, and M. Hubert, "Phase and amplitude-based clustering for functional data," *Computational Statistics & Data Analysis*, vol. 56, no. 7, pp. 2360–2374, 2012.

[34] R. De O Schmidt, R. Sadre, N. Melnikov, J. Schönwälder, and A. Pras, "Linking network usage patterns to traffic gaussianity fit," in *Networking Conference, 2014 IFIP*, June 2014, pp. 1–9.

[35] F. Simmross-Wattenberg, J. Asensio-Pérez, P. Casaseca-de-la Higuera, M. Martín-Fernández, I. Dimitriadis, and C. Alberola-López, "Anomaly detection in network traffic based on statistical inference and alpha-stable modeling," *Dependable and Secure Computing, IEEE Transactions on*, vol. 8, no. 4, pp. 494–509, July 2011.

[36] M.-S. Kim, Y. J. Won, and J. W. Hong, "Characteristic analysis of internet traffic from the perspective of flows," *Computer Communications*, vol. 29, no. 10, pp. 1639–1652, 2006, monitoring and Measurements of IP Networks. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140366405002768

[37] S. Saad, I. Traore, A. Ghorbani, B. Sayed, D. Zhao, W. Lu, J. Felix, and P. Hakimian, "Detecting P2P botnets through network behavior analysis and machine learning," in *Privacy, Security and Trust (PST), 2011 Ninth Annual International Conference on*, July 2011, pp. 174–180.

[38] C. Estan and G. Varghese, "New directions in traffic measurement and accounting," in *ACM SIGCOMM*, ser. SIGCOMM '02. New York, NY, USA: ACM, 2002, pp. 323–336. [Online]. Available: http://doi.acm.org/10.1145/633025.633056

[39] N. Brownlee and K. Claffy, "Understanding internet traffic streams: dragonflies and tortoises," *IEEE Communications Magazine*, vol. 40, no. 10, pp. 110–117, Oct 2002.

[40] A. Soule, K. Salamatia, N. Taft, R. Emilion, and K. Papagiannaki, "Flow classification by histograms: or how to go on safari in the internet," *ACM SIGMETRICS Performance Evaluation Review*, vol. 32, no. 1, pp. 49–60, 2004.

[41] V. Moreno, J. Ramos, P. Santiago del Rio, J. Garcia-Dorado, F. Gomez-Arribas, and J. Aracil, "Commodity packet capture engines: tutorial, cookbook and applicability," *Communications Surveys Tutorials, IEEE*, vol. PP, no. 99, pp. 1–1, 2015.

[42] D. Simoncelli, M. Dusi, F. Gringoli, and S. Niccolini, "Stream-monitoring with BlockMon: convergence of network measurements and data analytics platforms," *SIGCOMM Comput. Commun. Rev.*, vol. 43, p. 29–36, 2013. [Online]. Available: http://doi.acm.org/10.1145/2479957.2479962

[43] A. Papadogiannakis, M. Polychronakis, and E. P. Markatos, "Scap: Stream-oriented network traffic capture and analysis for high-speed networks," in *Proceedings of the 2013 Conference on Internet Measurement Conference*, ser. IMC '13. New York, NY, USA: ACM, 2013, pp. 441–454. [Online]. Available: http://doi.acm.org/10.1145/2504730.2504750

[44] A. Bar, P. Casas, L. Golab, and A. Finamore, "Dbstream: An online aggregation, filtering and processing system for network traffic monitoring," in *Wireless Communications and Mobile Computing Conference (IWCMC), 2014 International*, Aug 2014, pp. 611–616.

[45] Y. Lee, W. Kang, and Y. Lee, "A hadoop-based packet trace processing tool," in *Traffic Monitoring and Analysis*, ser. Lecture Notes in Computer Science, J. Domingo-Pascual, Y. Shavitt, and S. Uhlig, Eds. Springer Berlin Heidelberg, 2011, vol. 6613, pp. 51–63. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-20305-3_5

[46] Y. Lee and Y. Lee, "Toward scalable internet traffic measurement and analysis with hadoop," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 1, pp. 5–13, 2013.

[47] V. Moreno, P. M. Santiago del Río, J. Ramos, D. Muelas, J. L. García-Dorado, F. J. Gómez-Arribas, and J. Aracil, "Multi-granular, multi-purpose and multi-Gb/s monitoring on off-the-shelf systems," *International Journal of Network Management*, vol. 24, no. 4, pp. 221–234, 2014. [Online]. Available: http://dx.doi.org/10.1002/nem.1861

[48] J. L. García-Dorado, P. M. Santiago del Río, J. Ramos, D. Muelas, V. Moreno, J. E. López de Vergara, and J. Aracil, "Low-cost and high-performance: Voip monitoring and full-data retention at multi-gb/s rates using commodity hardware," *International Journal of Network Management*, vol. 24, no. 3, pp. 181–199, 2014. [Online]. Available: http://dx.doi.org/10.1002/nem.1858

[49] K. Xu, F. Wang, and H. Wang, "Lightweight and Informative Traffic Metrics for Data Center Monitoring," *Journal of Network and Systems Management*, vol. 20, no. 2, pp. 226–243, 2012. [Online]. Available: http://dx.doi.org/10.1007/s10922-011-9200-6

[50] J. L. García-Dorado, J. A. Hernández, J. Aracil, J. E. López de Vergara, and S. López-Buedo, "Characterization of the busy-hour traffic of IP networks based on their intrinsic features," *Computer Networks*, vol. 55, no. 9, pp. 2111 – 2125, 2011.

[51] T.-E. Wei, C.-H. Mao, A. Jeng, H.-M. Lee, H.-T. Wang, and D.-J. Wu, "Android malware detection via a latent network behavior analysis," in *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*, June 2012, pp. 1251–1258.

[52] F. Mata, J. L. García-Dorado, and J. Aracil, "Detection of traffic changes in large-scale backbone networks: The case of the Spanish academic network," *Computer Networks*, vol. 56, no. 2, pp. 686 – 702, 2012.

[53] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft, "Structural analysis of network traffic flows," *SIGMETRICS Perform. Eval. Rev.*, vol. 32, no. 1, pp. 61–72, Jun. 2004. [Online]. Available: http://doi.acm.org/10.1145/1012888.1005697

[54] J. L. García-Dorado, J. Aracil, J. A. Hernández, and J. E. López de Vergara, "A queueing equivalent thresholding method for thinning traffic captures," in *Network Operations and Management Symposium, 2008. NOMS 2008. IEEE*, April 2008, pp. 176–183.

[55] K. Kyriakopoulos and D. Parish, "A live system for wavelet compression of high speed computer network measurements," in *Passive and Active Network Measurement*, ser. Lecture Notes in Computer Science, S. Uhlig, K. Papagiannaki,

and O. Bonaventure, Eds. Springer Berlin Heidelberg, 2007, vol. 4427, pp. 241–244. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-71617-4_27

[56] V. Menkovski, A. Oredope, A. Liotta, and A. Cuadra Sánchez, "Optimized online learning for QoE prediction," in *Proceedings 21st Benelux Conference on Artificial Intelligence. BNAIC'09*, October 2009, pp. 169–176.

[57] M. Nassar, O. Dabbebi, R. Badonnel, and O. Festor, "Risk management in VoIP infrastructures using support vector machines," in *Network and Service Management (CNSM), 2010 International Conference on*, Oct 2010, pp. 48–55.

[58] B. Li, J. Springer, G. Bebis, and M. H. Gunes, "A survey of network flow applications," *Journal of Network and Computer Applications*, vol. 36, no. 2, pp. 567–581, 2013.

[59] A. Sperotto, G. Schaffrath, R. Sadre, C. Morariu, A. Pras, and B. Stiller, "An overview of IP flow-based intrusion detection," *Communications Surveys Tutorials, IEEE*, vol. 12, no. 3, pp. 343–356, Third 2010.

[60] J. Rejchrt, T. Jirsik, and J. Vykopal, "Time Series Solver," Masarykova univerzita, 2013. [Online]. Available: http://www.muni.cz/ics/services/csirt/tools/tss

[61] M. Febrero-Bande and M. Oviedo de la Fuente, "Statistical computing in functional data analysis: The R package fda.usc," *Journal of Statistical Software*, vol. 51, no. 4, pp. 1–28, 2012. [Online]. Available: http://www.jstatsoft.org/v51/i04/

[62] W. Xiong, H. Hu, N. Xiong, L. T. Yang, W.-C. Peng, X. Wang, and Y. Qu, "Anomaly secure detection methods by analyzing dynamic characteristics of the network traffic in cloud communications," *Information Sciences*, vol. 258, no. 0, pp. 403 – 415, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025513002971

[63] A. Aguado, V. López, J. Marhuenda, J.-P. Fernández-Palacios *et al.*, "ABNO: a feasible SDN approach for multi-vendor IP and optical networks," in *Optical Fiber Communication Conference*. Optical Society of America, 2014, pp. Th3I–5.

[64] J. E. Angus, "The probability integral transform and related results," *SIAM Review*, vol. 36, no. 4, pp. 652–654, 1994. [Online]. Available: http://dx.doi.org/10.1137/1036146

[65] J. Williams, K. Sochats, and E. Morse, "Visualization," *Annual review of information science and technology*, vol. 30, pp. 161–207, 1995.

[66] J. A. Wellner *et al.*, "A glivenko-cantelli theorem and strong laws of large numbers for functions of order statistics," *The Annals of Statistics*, vol. 5, no. 3, pp. 473–480, 1977.

[67] M. Gordo, "Detección forense de ataques usando trazas de red," B.S. Thesis, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 2015.

[68] J. L. García-Dorado, J. A. Hernández, J. Aracil, J. E. López de Vergara, F. Montserrat, E. Robles, and T. de Miguel, "On the duration and spatial characteristics of internet traffic measurement experiments," *IEEE Communications Magazine*, vol. 46, no. 11, pp. 152–154, 2008.

[69] E. S. Gardner, "Exponential smoothing: The state of the art – part II," *International journal of forecasting*, vol. 22, no. 4, pp. 637–666, 2006.

[70] V. Jacobson, "Congestion avoidance and control," *ACM SIGCOMM Computer Communication Review*, vol. 18, no. 4, pp. 314–329, 1988.

LISTS

List of equations

List of figures

List of tables

Acronyms

ABNO	Application-based Network Operations
ACM	Association for Computing Machinery
CDF	Cumulative Distribution Function
DDoS	Distributed Denial of Service
DPI	Deep Packet Inspection
DSW	Data Stream Warehousing
ECDF	Empirical Cumulative Distribution Function
FCAPS	Fault, Configuration, Accounting, Performance, Security
FDA	Functional Data Analysis
FPCA	Functional Principal Component Analysis
IPFIX	IP Flow Information eXport
NTMA	Network Traffic Monitoring and Analysis
PCA	Principal Component Analysis
QoE	Quality of Experience
QoS	Quality of Service
SDN	Software Defined Networking
SNMP	Simple Network Management Protocol
VoIP	Voice over IP
w.l.o.g.	without loss of generality

TERMINOLOGY

Banach space *Complete normed vector space.*

Borel set *Any set in a topological space that can be formed from open sets through the operations of countable union, countable intersection, and relative complement.*

Systems management functional area *A category of systems management user requirements.*

Hadoop *Framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.*

Hilbert space *Vector space with an inner product.*

Indicator function *Functions defined on a set S, with value 1 in the elements of a certain subset A, and 0 in the elements of the complementary of A.*

Karhunen-Loeve expansion *Functional extension of the principal component analysis to stochastic processes.*

MapReduce *MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key.*

Probability space *Mathematical structure that links a sample space to a set of possible events and to a function that assigns probabilities to the events.*

Simple function *Functions defined as linear combinations of indicator functions.*